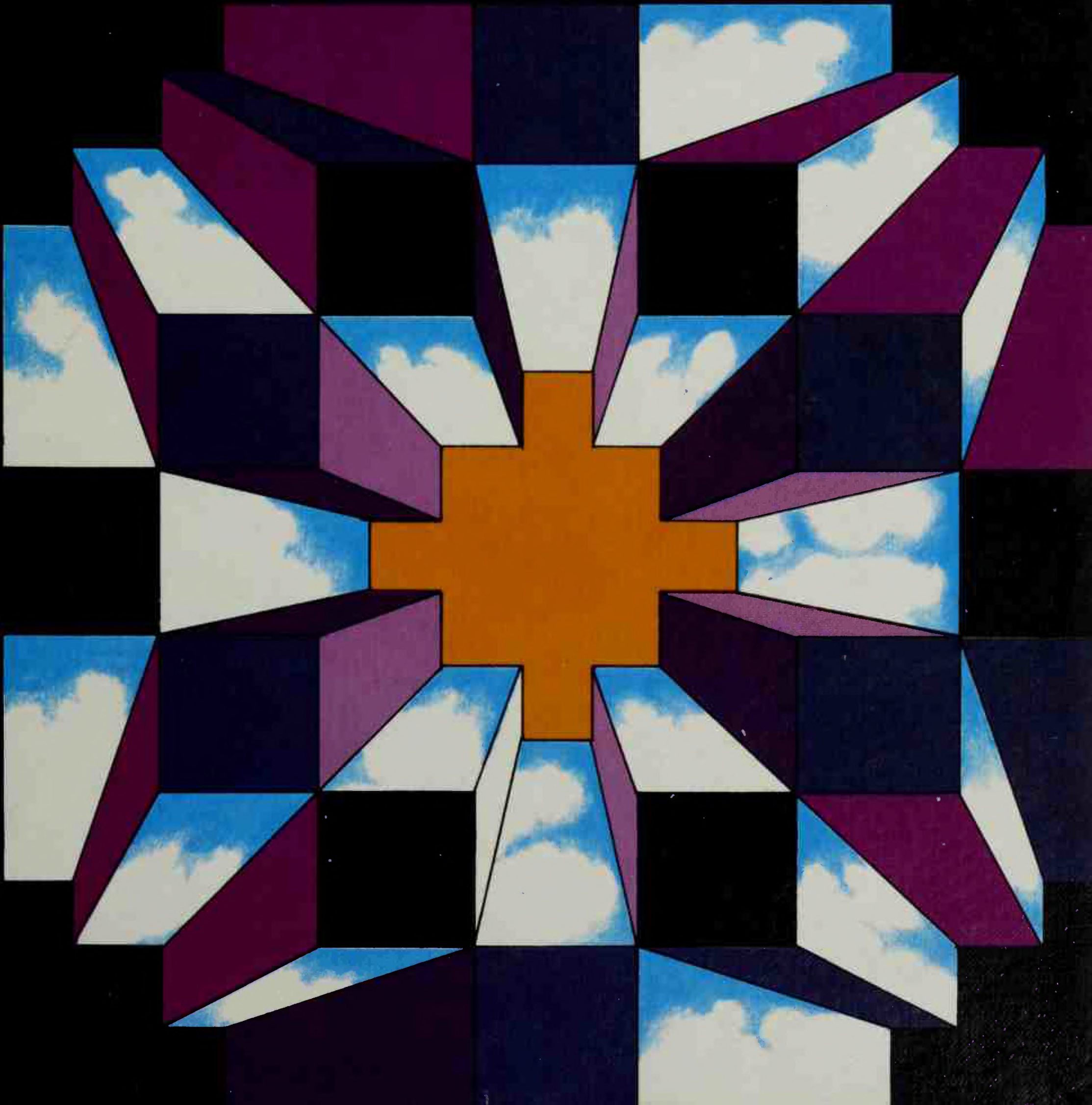
PSYCHOLOGICAL TESTING

Fourth Edition

Anne Anastasi



SHOUT
SHARE
IT

Psychological Testing

ANNE ANASTASI
Professor of Psychology, Fordham University

Psychological Testing

FOURTH EDITION

MACMILLAN PUBLISHING CO., INC.
New York

Collier Macmillan Publishers
London

COPYRIGHT © 1976, ANNE ANASTASI

PRINTED IN THE UNITED STATES OF AMERICA

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the Publisher.

Earlier editions copyright 1954 and © 1961 by Macmillan Publishing Co., Inc., and copyright © 1968 by Anne Anastasi.

MACMILLAN PUBLISHING CO., INC.
866 Third Avenue, New York, New York 10022

COLLIER MACMILLAN CANADA, LTD.

Library of Congress Cataloging in Publication Data

Anastasi, Anne, (date)
Psychological testing.

Bibliography: p.
Includes indexes.
1. Mental tests. 2. Personality tests. I. Title.
[DNLM: 1. Psychological tests. WM145 A534p]
BF431.A573 1976 153.9 75-2206
ISBN 0-02-302980-3 (Hardbound)
ISBN 0-02-979110-3 (International Edition)

Printing: 7 8 Year: 9 0 1 2

PREFACE

IN A revised edition, one expects both similarities and differences. This edition shares with the earlier versions the objectives and basic approach of the book. The primary goal of this text is still to contribute toward the proper evaluation of psychological tests and the correct interpretation and use of test results. This goal calls for several kinds of information: (1) an understanding of the major principles of test construction, (2) psychological knowledge about the behavior being assessed, (3) sensitivity to the social and ethical implications of test use, and (4) broad familiarity with the types of available instruments and the sources of information about tests. A minor innovation in the fourth edition is the addition of a suggested outline for test evaluation (Appendix C).

In successive editions, it has been necessary to exercise more and more restraint to keep the number of specific tests discussed in the book from growing with the field—it has never been my intention to provide a miniature *Mental Measurements Yearbook!* Nevertheless, I am aware that principles of test construction and interpretation can be better understood when applied to particular tests. Moreover, acquaintance with the major types of available tests, together with an understanding of their special contributions and limitations, is an essential component of knowledge about contemporary testing. For these reasons, specific tests are again examined and evaluated in Parts 3, 4, and 5. These tests have been chosen either because they are outstanding examples with which the student of testing should be familiar or because they illustrate some special point of test construction or interpretation. In the text itself, the principal focus is on types of tests rather than on specific instruments. At the same time, Appendix E contains a classified list of over 250 tests, including not only those cited in the text but also others added to provide a more representative sample.

As for the differences—they loomed especially large during the preparation of this edition. Much that has happened in human society since the mid-1960's has had an impact on psychological testing. Some of these developments were briefly described in the last two chapters of the third edition. Today they have become part of the mainstream of psychological testing and have been accordingly incorporated in the appropriate sections throughout the book. Recent changes in psychological testing that are reflected in the present edition can be described on three levels: (1) general orientation toward testing, (2) substantive and methodological developments, and (3) "ordinary progress" such as the publication of new tests and revision of earlier tests.

An example of changes on the first level is the increasing awareness of the ethical, social, and legal implications of testing. In the present edition, this topic has been expanded and treated in a separate chapter early in the book (Ch. 3) and in Appendixes A and B. A cluster of related developments represents a broadening of test uses. Besides the traditional applications of tests in selection and diagnosis, increasing attention is being given to administering tests for self-knowledge and self-development, and to training individuals in the use of their own test results in decision making (Chs. 3 and 4). In the same category are the continuing replacement of global scores with multitrait profiles and the application of classification strategies, whereby "everyone can be above average" in one or more socially valued variables (Ch. 7). From another angle, efforts are being made to modify traditional interpretations of test scores, in both cognitive and noncognitive areas, in the light of accumulating psychological knowledge. In this edition, Chapter 12 brings together psychological issues in the interpretation of intelligence test scores, touching on such problems as stability and change in intellectual level over time; the nature of intelligence; and the testing of intelligence in early childhood, in old age, and in different cultures. Another example is provided by the increasing emphasis on situational specificity and person-by-situation interactions in personality testing, stimulated in large part by the social-learning theorists (Ch. 17).

The second level, covering substantive and methodological changes, is illustrated by the impact of computers on the development, administration, scoring, and interpretation of tests (see especially Chs. 4, 11, 13, 17, 18, 19). The use of computers in administering or managing instructional programs has also stimulated the development of criterion-referenced tests, although other conditions have contributed to the upsurge of interest in such tests in education. Criterion-referenced tests are discussed principally in Chapters 4, 5, and 14. Other types of instruments that have risen to prominence and have received fuller treatment in the present edition include: tests for identifying specific learning disabilities (Ch. 16), inventories and other devices for use in behavior modification programs (Ch. 20), instruments for assessing early childhood education (Ch. 14), Piagetian "ordinal" scales (Chs. 10 and 14), basic education and literacy tests for adults (Chs. 13 and 14), and techniques for the assessment of environments (Ch. 20). Problems to be considered in the assessment of minority groups, including the question of test bias, are examined from different angles in Chapters 3, 7, 8, and 12.

At the third level, it may be noted that over 100 of the tests listed in this edition have been either initially published or revised since the publication of the preceding edition (1968). Major examples include the McCarthy Scales of Children's Abilities, the WISC-R, the 1972 Stanford-Binet norms (with all the resulting readjustments in interpretations), Forms S and T of the DAT (including a computerized Career Planning Program), the Strong-Campbell Interest Inventory (merged form of the SVIB), and the latest revisions of the Stanford Achievement Test and the Metropolitan Readiness Tests.

It is a pleasure to acknowledge the assistance received from many sources in the preparation of this edition. The completion of the project was facilitated by a one-semester Faculty Fellowship awarded by Fordham University and by a grant from the Fordham University Research Council covering principally the services of a research assistant. These services were performed by Stanley Friedland with an unusual combination of expertise, responsibility, and graciousness. I am indebted to the many authors and test publishers who provided reprints, unpublished manuscripts, specimen sets of tests, and answers to my innumerable inquiries by mail and telephone. For assistance extending far beyond the interests and responsibilities of any single publisher, I am especially grateful to Anna Dragositz of Educational Testing Service and Blythe Mitchell of Harcourt Brace Jovanovich, Inc. I want to acknowledge the significant contribution of John T. Cowles of the University of Pittsburgh, who assumed complete responsibility for the preparation of the *Instructor's Manual* to accompany this text.

For informative discussions and critical comments on particular topics, I want to convey my sincere thanks to William H. Angoff of Educational Testing Service and to several members of the Fordham University Psychology Department, including David R. Chabot, Marvin Reznikoff, Reuben M. Schonebaum, and Warren W. Tryon. Grateful acknowledgment is also made of the thoughtful recommendations submitted by course instructors in response to the questionnaire distributed to current users of the third edition. Special thanks in this connection are due to Mary Carol Cahill for her extensive, constructive, and wide-ranging suggestions. I wish to express my appreciation to Victoria Overton of the Fordham University library staff for her efficient and courteous assistance in bibliographic matters. Finally, I am happy to record the contributions of my husband, John Porter Foley, Jr., who again participated in the solution of countless problems at all stages in the preparation of the book.

A. A.

CONTENTS

PART 1
CONTEXT OF PSYCHOLOGICAL TESTING

1. FUNCTIONS AND ORIGINS OF PSYCHOLOGICAL TESTING 3

Current uses of psychological tests 3
Early interest in classification and training of the mentally retarded 5
The first experimental psychologists 7
Contributions of Francis Galton 8
Cattell and the early "mental tests" 9
Binet and the rise of intelligence tests 10
Group testing 12
Aptitude testing 13
Standardized achievement tests 16
Measurement of personality 18
Sources of information about tests 20

2. NATURE AND USE OF PSYCHOLOGICAL TESTS 23

What is a psychological test? 23
Reasons for controlling the use of psychological tests 30
Test administration 32
Rapport 34
Test anxiety 37
Examiner and situational variables 39
Coaching, practice, and test sophistication 41

3. SOCIAL AND ETHICAL IMPLICATIONS OF TESTING 45

User qualifications 45
Testing instruments and procedures 47
Protection of privacy 49
Confidentiality 52
Communicating test results 56
Testing and the civil rights of minorities 57

PART 2
PRINCIPLES OF PSYCHOLOGICAL TESTING

4. NORMS AND THE INTERPRETATION OF TEST SCORES 67

Statistical concepts 68
Developmental norms 73
Within-group norms 77
Relativity of norms 88
Computer utilization in the interpretation of test scores 94
Criterion-referenced testing 96

5. RELIABILITY 103

The correlation coefficient 104
Types of reliability 110
Reliability of speeded tests 122
Dependence of reliability coefficients on the sample tested 125
Standard error of measurement 127
Reliability of criterion-referenced tests 131

6. VALIDITY: BASIC CONCEPTS 134

Content validity 134
Criterion-related validity 140
Construct validity 151
Overview 158

7. VALIDITY: MEASUREMENT AND INTERPRETATION 162

Validity coefficient and error of estimate 163
Test validity and decision theory 167
Moderator variables 177
Combining information from different tests 180
Use of tests for classification decisions 186
Statistical analyses of test bias 191

8. ITEM ANALYSIS 198

Item difficulty 199
Item validity 206
Internal consistency 215
Item analysis of speeded tests 217
Cross validation 219
Item-group interaction 222

PART 3

TESTS OF GENERAL INTELLECTUAL LEVEL

9. INDIVIDUAL TESTS 229

Stanford-Binet Intelligence Scale 230
Wechsler Adult Intelligence Scale 245
Wechsler Intelligence Scale for Children 255
Wechsler Preschool and Primary Scale of Intelligence 260

10. TESTS FOR SPECIAL POPULATIONS 265

Infant and preschool testing 266
Testing the physically handicapped 281
Cross-cultural testing 287

11. GROUP TESTING 299

Group tests versus individual tests 299
Multilevel batteries 305
Tests for the college level and beyond 318

12. PSYCHOLOGICAL ISSUES IN INTELLIGENCE TESTING 326

Longitudinal studies of intelligence 327
Intelligence in early childhood 332
Problems in the testing of adult intelligence 337
Problems in cross-cultural testing 343
Nature of intelligence 349

PART 4

TESTS OF SEPARATE ABILITIES

13. MEASURING MULTIPLE APTITUDES 361

Factor analysis 362
Theories of trait organization 369
Multiple aptitude batteries 378
Measurement of creativity 388

14. EDUCATIONAL TESTING 398

Achievement tests: their nature and uses 398
General achievement batteries 403
Standardized tests in separate subjects 410
Teacher-made classroom tests 412

Diagnostic and criterion-referenced tests 417
Specialized prognostic tests 423
Assessment in early childhood education 425

15. OCCUPATIONAL TESTING 435

Validation of industrial tests 435
Short screening tests for industrial personnel 439
Special aptitude tests 442
Testing in the professions 458

16. CLINICAL TESTING 464

Diagnostic use of intelligence tests 465
Special tests for detecting cognitive dysfunction 471
Identifying specific learning disabilities 478
Clinical judgment 482
Report writing 487

PART 5
PERSONALITY TESTS

17. SELF-REPORT INVENTORIES 493

Content validation 494
Empirical criterion keying 496
Factor analysis in test development 506
Personality theory in test development 510
Test-taking attitudes and response sets 515
Situational specificity 521
Evaluation of personality inventories 524

18. MEASURES OF INTERESTS, ATTITUDES, AND VALUES 527

Interest inventories 528
Opinion and attitude measurement 543
Attitude scales 546
Assessment of values and related variables 552

19. PROJECTIVE TECHNIQUES 558

Nature of projective techniques 558
Inkblot techniques 559
Thematic Apperception Test and related instruments 565
Other projective techniques 569
Evaluation of projective techniques 576

20. OTHER ASSESSMENT TECHNIQUES 588
"Objective" performance tests 588
Situational tests 593
Self-concepts and personal constructs 598
Assessment techniques in behavior modification programs 604
Observer reports 606
Biographical inventories 614
The assessment of environments 616

REFERENCES 621

APPENDIXES **681**

A. Ethical Standards of Psychologists 682

B. Guidelines on Employee Selection Procedures (EEOC) 691

Guidelines for Reporting Criterion-Related and Content Validity (OFCC) 699

C. A Suggested Outline for Test Evaluation 705

D. Test Publishers 707

E. Classified List of Representative Tests 708

Author Index 723

Subject Index 737

PART 1

Context of Psychological Testing

CHAPTER 1

Functions and Origins of Psychological Testing

ANYONE reading this book today could undoubtedly illustrate what is meant by a psychological test. It would be easy enough to recall a test the reader himself has taken in school, in college, in the armed services, in the counseling center, or in the personnel office. Or perhaps the reader has served as a subject in an experiment in which standardized tests were employed. This would certainly not have been the case fifty years ago. Psychological testing is a relatively young branch of one of the youngest of the sciences.

CURRENT USES OF PSYCHOLOGICAL TESTS

Basically, the function of psychological tests is to measure differences between individuals or between the reactions of the same individual on different occasions. One of the first problems that stimulated the development of psychological tests was the identification of the mentally retarded. To this day, the detection of intellectual deficiencies remains an important application of certain types of psychological tests. Related clinical uses of tests include the examination of the emotionally disturbed, the delinquent, and other types of behavioral deviants. A strong impetus to the early development of tests was likewise provided by problems arising in education. At present, schools are among the largest test users. The classification of children with reference to their ability to profit from different types of school instruction, the identification of the intellectually retarded on the one hand and the gifted on the other, the diagnosis of academic failures, the educational and vocational counseling of high school and college students, and the selection of applicants for professional and other special schools are among the many educational uses of tests.

The selection and classification of industrial personnel represent another major application of psychological testing. From the assembly-line operator or filing clerk to top management, there is scarcely a type of job for which some kind of psychological test has not proved helpful in such matters as hiring, job assignment, transfer, promotion, or termination. To be sure, the effective employment of tests in many of these situations, especially in connection with high-level jobs, usually requires that the tests be used as an adjunct to skillful interviewing, so that test scores may be properly interpreted in the light of other background information about the individual. Nevertheless, testing constitutes an important part of the total personnel program. A closely related application of psychological testing is to be found in the selection and classification of military personnel. From simple beginnings in World War I, the scope and variety of psychological tests employed in military situations underwent a phenomenal increase during World War II. Subsequently, research on test development has been continuing on a large scale in all branches of the armed services.

The use of tests in counseling has gradually broadened from a narrowly defined guidance regarding educational and vocational plans to an involvement with all aspects of the person's life. Emotional well-being and effective interpersonal relations have become increasingly prominent objectives of counseling. There is growing emphasis, too, on the use of tests to enhance self-understanding and personal development. Within this framework, test scores are part of the information given to the individual as aids to his own decision-making processes.

It is clearly evident that psychological tests are currently being employed in the solution of a wide range of practical problems. One should not, however, lose sight of the fact that such tests are also serving important functions in basic research. Nearly all problems in differential psychology, for example, require testing procedures as a means of gathering data. As illustrations, reference may be made to studies on the nature and extent of individual differences, the identification of psychological traits, the measurement of group differences, and the investigation of biological and cultural factors associated with behavioral differences. For all such areas of research—and for many others—the precise measurement of individual differences made possible by well-constructed tests is an essential prerequisite. Similarly, psychological tests provide standardized tools for investigating such varied problems as life-span developmental changes within the individual, the relative effectiveness of different educational procedures, the outcomes of psychotherapy, the impact of community programs, and the influence of noise on performance.

From the many different uses of psychological tests, it follows that some knowledge of such tests is needed for an adequate understanding of most fields of contemporary psychology. It is primarily with this end in view that the present book has been prepared. The book is not designed to make the individual either a skilled examiner and test administrator or an expert on test construction. It is directed, not to the test specialist, but to the general student of psychology. Some acquaintance with the leading current tests is necessary in order to understand references to the use of such tests in the psychological literature. And a proper evaluation and interpretation of test results must ultimately rest on a knowledge of how the tests were constructed, what they can be expected to accomplish, and what are their peculiar limitations. Today a familiarity with tests is required, not only by those who give or construct tests, but by the general psychologist as well.

A brief overview of the historical antecedents and origins of psychological testing will provide perspective and should aid in the understanding of present-day tests.[1] The direction in which contemporary psychological testing has been progressing can be clarified when considered in the light of the precursors of such tests. The special limitations as well as the advantages that characterize current tests likewise become more intelligible when viewed against the background in which they originated.

The roots of testing are lost in antiquity. DuBois (1966) gives a provocative and entertaining account of the system of civil service examinations prevailing in the Chinese empire for some three thousand years. Among the ancient Greeks, testing was an established adjunct to the educational process. Tests were used to assess the mastery of physical as well as intellectual skills. The Socratic method of teaching, with its interweaving of testing and teaching, has much in common with today's programed learning. From their beginnings in the middle ages, European universities relied on formal examinations in awarding degrees and honors. To identify the major developments that shaped contemporary testing, however, we need go no farther than the nineteenth century. It is to these developments that we now turn.

EARLY INTEREST IN CLASSIFICATION AND TRAINING OF THE MENTALLY RETARDED

The nineteenth century witnessed a strong awakening of interest in the humane treatment of the mentally retarded and the insane. Prior to that time, neglect, ridicule, and even torture had been the common lot of these unfortunates. With the growing concern for the proper care of mental deviates came a realization that some uniform criteria for identifying and classifying these cases were required. The establishment of many special institutions for the care of the mentally retarded in both Europe and America made the need for setting up admission standards and an objective system of classification especially urgent. First it was necessary to differentiate between the insane and the mentally retarded. The former manifested emotional disorders that might or might not be accompanied by intellectual deterioration from an initially normal level; the latter were characterized essentially by intellectual defect that had been present from birth or early infancy. What is probably the first explicit statement of this distinction is to be found in a two-volume work published in 1838 by the French physician Esquirol (1838), in which over one hundred pages are devoted to mental retardation. Esquirol also pointed out that there are many degrees of mental retardation, varying along a continuum from normality to low-grade idiocy. In the effort to develop some system for classifying the different degrees and varieties of retardation, Esquirol tried several procedures but concluded that the individual's use of language provides the most dependable criterion of his intellectual level. It is interesting to note that current criteria of mental retardation are also largely linguistic and that present-day intelligence tests are heavily loaded with verbal content. The important part verbal ability plays in our concept of intelligence will be repeatedly demonstrated in subsequent chapters.

---

[1] A more detailed account of the early origins of psychological tests can be found in Goodenough (1949) and J. Peterson (1926). See also Boring (1950) and Murphy and Kovach (1972) for more general background, DuBois (1970) for a brief but comprehensive history of psychological testing, and Anastasi (1965) for historical antecedents of the study of individual differences.

Of special significance are the contributions of another French physician, Seguin, who pioneered in the training of the mentally retarded. Having rejected the prevalent notion of the incurability of mental retardation, Seguin (1866) experimented for many years with what he termed the physiological method of training; and in 1837 he established the first school devoted to the education of mentally retarded children. In 1848 he emigrated to America, where his ideas gained wide recognition. Many of the sense-training and muscle-training techniques currently in use in institutions for the mentally retarded were originated by Seguin. By these methods, severely retarded children are given intensive exercise in sensory discrimination and in the development of motor control. Some of the procedures developed by Seguin for this purpose were eventually incorporated into performance or nonverbal tests of intelligence. An example is the Seguin Form Board, in which the individual is required to insert variously shaped blocks into the corresponding recesses as quickly as possible.

More than half a century after the work of Esquirol and Seguin, the French psychologist Alfred Binet urged that children who failed to respond to normal schooling be examined before dismissal and, if considered educable, be assigned to special classes (T. H. Wolf, 1973). With his fellow members of the Society for the Psychological Study of the Child, Binet stimulated the Ministry of Public Instruction to take steps to improve the condition of retarded children. A specific outcome was the establishment of a ministerial commission for the study of retarded children, to which Binet was appointed. This appointment was a momentous event in the history of psychological testing, of which more will be said later.

THE FIRST EXPERIMENTAL PSYCHOLOGISTS

The early experimental psychologists of the nineteenth century were not, in general, concerned with the measurement of individual differences. The principal aim of psychologists of that period was the formulation of generalized descriptions of human behavior. It was the uniformities rather than the differences in behavior that were the focus of attention. Individual differences were either ignored or were accepted as a necessary evil that limited the applicability of the generalizations. Thus, the fact that one individual reacted differently from another when observed under identical conditions was regarded as a form of error. The presence of such error, or individual variability, rendered the generalizations approximate rather than exact. This was the attitude toward individual differences that prevailed in such laboratories as that founded by Wundt at Leipzig in 1879, where many of the early experimental psychologists received their training.

In their choice of topics, as in many other phases of their work, the founders of experimental psychology reflected the influence of their backgrounds in physiology and physics. The problems studied in their laboratories were concerned largely with sensitivity to visual, auditory, and other sensory stimuli and with simple reaction time. This emphasis on sensory phenomena was in turn reflected in the nature of the first psychological tests, as will be apparent in subsequent sections.

Still another way in which nineteenth-century experimental psychology influenced the course of the testing movement may be noted. The early psychological experiments brought out the need for rigorous control of the conditions under which observations were made. For example, the wording of directions given to the subject in a reaction-time experiment might appreciably increase or decrease the speed of the subject's response. Or again, the brightness or color of the surrounding field could markedly alter the appearance of a visual stimulus. The importance of making observations on all subjects under standardized conditions was thus vividly demonstrated. Such standardization of procedure eventually became one of the special earmarks of psychological tests.

CONTRIBUTIONS OF FRANCIS GALTON

It was the English biologist Sir Francis Galton who was primarily responsible for launching the testing movement. A unifying factor in Galton's numerous and varied research activities was his interest in human heredity. In the course of his investigations on heredity, Galton realized the need for measuring the characteristics of related and unrelated persons. Only in this way could he discover, for example, the exact degree of resemblance between parents and offspring, brothers and sisters, cousins, or twins. With this end in view, Galton was instrumental in inducing a number of educational institutions to keep systematic anthropometric records on their students. He also set up an anthropometric laboratory at the International Exposition of 1884 where, by paying threepence, visitors could be measured in certain physical traits and could take tests of keenness of vision and hearing, muscular strength, reaction time, and other simple sensorimotor functions. When the exposition closed, the laboratory was transferred to South Kensington Museum, London, where it operated for six years. By such methods, the first large, systematic body of data on individual differences in simple psychological processes was gradually accumulated.

Galton himself devised most of the simple tests administered at his anthropometric laboratory, many of which are still familiar either in their original or in modified forms. Examples include the Galton bar for visual discrimination of length, the Galton whistle for determining the highest audible pitch, and graduated series of weights for measuring kinesthetic discrimination. It was Galton's belief that tests of sensory discrimination could serve as a means of gauging a person's intellect. In this respect, he was partly influenced by the theories of Locke. Thus Galton wrote: "The only information that reaches us concerning outward events appears to pass through the avenue of our senses; and the more perceptive the senses are of difference, the larger is the field upon which our judgment and intelligence can act" (Galton, 1883, p. 27). Galton had also noted that idiots tend to be defective in the ability to discriminate heat, cold, and pain—an observation that further strengthened his conviction that sensory discriminative capacity "would on the whole be highest among the intellectually ablest" (Galton, 1883, p. 29).

Galton also pioneered in the application of rating-scale and questionnaire methods, as well as in the use of the free association technique subsequently employed for a wide variety of purposes. A further contribution of Galton is to be found in his development of statistical methods for the analysis of data on individual differences. Galton selected and adapted a number of techniques previously derived by mathematicians. These techniques he put in such form as to permit their use by the mathematically untrained investigator who might wish to treat test results quantitatively. He thereby extended enormously the application of statistical procedures to the analysis of test data. This phase of Galton's work has been carried forward by many of his students, the most eminent of whom was Karl Pearson.

CATTELL AND THE EARLY "MENTAL TESTS"

An especially prominent position in the development of psychological testing is occupied by the American psychologist James McKeen Cattell. The newly established science of experimental psychology and the still newer testing movement merged in Cattell's work. For his doctorate at Leipzig, he completed a dissertation on individual differences in reaction time, despite Wundt's resistance to this type of investigation. While lecturing at Cambridge in 1888, Cattell's own interest in the measurement of individual differences was reinforced by contact with Galton. On his return to America, Cattell was active both in the establishment of laboratories for experimental psychology and in the spread of the testing movement.

In an article written by Cattell in 1890, the term "mental test" was used for the first time in the psychological literature. This article described a series of tests that were being administered annually to college students in the effort to determine their intellectual level. The tests, which had to be administered individually, included measures of muscular strength, speed of movement, sensitivity to pain, keenness of vision and of hearing, weight discrimination, reaction time, memory, and the like. In his choice of tests, Cattell shared Galton's view that a measure of intellectual functions could be obtained through tests of sensory discrimination and reaction time. Cattell's preference for such tests was also bolstered by the fact that simple functions could be measured with precision and accuracy, whereas the development of objective measures for the more complex functions seemed at that time a well-nigh hopeless task.

Cattell's tests were typical of those to be found in a number of test series developed during the last decade of the nineteenth century. Such test series were administered to schoolchildren, college students, and miscellaneous adults. At the Columbian Exposition held in Chicago in 1893, Jastrow set up an exhibit at which visitors were invited to take tests of sensory, motor, and simple perceptual processes and to compare their skill with the norms (J. Peterson, 1926; Philippe, 1894). A few attempts to evaluate such early tests yielded very discouraging results. The individual's performance showed little correspondence from one test to another (Sharp, 1898–1899; Wissler, 1901), and it exhibited little or no relation to independent estimates of intellectual level based on teachers' ratings (Bolton, 1891–1892; J. A. Gilbert, 1894) or academic grades (Wissler, 1901).

A number of test series assembled by European psychologists of the period tended to cover somewhat more complex functions. Kraepelin (1895), who was interested primarily in the clinical examination of psychiatric patients, prepared a long series of tests to measure what he regarded as basic factors in the characterization of an individual. The tests, employing chiefly simple arithmetic operations, were designed to measure practice effects, memory, and susceptibility to fatigue and to distraction. A few years earlier, Oehrn (1889), a pupil of Kraepelin, had employed tests of perception, memory, association, and motor functions in an investigation on the interrelations of psychological functions. Another German psychologist, Ebbinghaus (1897), administered tests of arithmetic computation, memory span, and sentence completion to schoolchildren. The most complex of the three tests, sentence completion, was the only one that showed a clear correspondence with the children's scholastic achievement.

Like Kraepelin, the Italian psychologist Ferrari and his students were interested primarily in the use of tests with pathological cases (Guicciardi & Ferrari, 1896). The test series they devised ranged from physiological measures and motor tests to apprehension span and the interpretation of pictures. In an article published in France in 1895, Binet and Henri criticized most of the available test series as being too largely sensory and as concentrating unduly on simple, specialized abilities. They argued further that, in the measurement of the more complex functions, great precision is not necessary, since individual differences are larger in these functions. An extensive and varied list of tests was proposed, covering such functions as memory, imagination, attention, comprehension, suggestibility, aesthetic appreciation, and many others. In these tests we can recognize the trends that were eventually to lead to the development of the famous Binet intelligence scales.

BINET AND THE RISE OF INTELLIGENCE TESTS

Binet and his co-workers devoted many years to active and ingenious research on ways of measuring intelligence. Many approaches were tried, including even the measurement of cranial, facial, and hand form, and the analysis of handwriting. The results, however, led to a growing conviction that the direct, even though crude, measurement of complex intellectual functions offered the greatest promise. Then a specific situation arose that brought Binet's efforts to immediate practical fruition. In 1904, the Minister of Public Instruction appointed Binet to the previously cited commission to study procedures for the education of retarded children. It was in connection with the objectives of this commission that Binet, in collaboration with Simon, prepared the first Binet-Simon Scale (Binet & Simon, 1905).

This scale, known as the 1905 scale, consisted of 30 problems or tests arranged in ascending order of difficulty. The difficulty level was determined empirically by administering the tests to 50 normal children aged 3 to 11 years, and to some mentally retarded children and adults. The tests were designed to cover a wide variety of functions, with special emphasis on judgment, comprehension, and reasoning, which Binet regarded as essential components of intelligence. Although sensory and perceptual tests were included, a much greater proportion of verbal content was found in this scale than in most test series of the time. The 1905 scale was presented as a preliminary and tentative instrument, and no precise objective method for arriving at a total score was formulated.

In the second, or 1908, scale, the number of tests was increased, some unsatisfactory tests from the earlier scale were eliminated, and all tests were grouped into age levels on the basis of the performance of about 300 normal children between the ages of 3 and 13 years. Thus, in the 3-year level were placed all tests passed by 80 to 90 percent of normal 3-year-olds; in the 4-year-level, all tests similarly passed by normal 4-year-olds; and so on to age 13. The child's score on the entire test could then be expressed as a *mental level* corresponding to the age of normal children whose performance he equaled. In the various translations and adaptations of the Binet scales, the term "mental age" was commonly substituted for "mental level." Since mental age is such a simple concept to grasp, the introduction of this term undoubtedly did much to popularize intelligence testing.[2] Binet himself, however, avoided the term "mental age" because of its unverified developmental implications and preferred the more neutral term "mental level" (T. H. Wolf, 1973).

A third revision of the Binet-Simon Scale appeared in 1911, the year of Binet's untimely death. In this scale, no fundamental changes were introduced. Minor revisions and relocations of specific tests were instituted. More tests were added at several year levels, and the scale was extended to the adult level.

Even prior to the 1908 revision, the Binet-Simon tests attracted wide

[2] Goodenough (1949, pp. 50–51) notes that in 1887, 21 years before the appearance of the 1908 Binet-Simon Scale, S. E. Chaille published in the *New Orleans Medical and Surgical Journal* a series of tests for infants arranged according to the age at which the tests are commonly passed. Partly because of the limited circulation of the journal and partly, perhaps, because the scientific community was not ready for it, the significance of this age-scale concept passed unnoticed at the time. Binet's own scale was influenced by the work of some of his contemporaries, notably Blin and Damaye, who prepared a set of oral questions from which they derived a single global score for each child (T. H. Wolf, 1973).

attention among psychologists throughout the world. Translations and adaptations appeared in many languages. In America, a number of different revisions were prepared, the most famous of which is the one developed under the direction of L. M. Terman at Stanford University, and known as the Stanford-Binet (Terman, 1916). It was in this test that the intelligence quotient (IQ), or ratio between mental age and chronological age, was first used. The latest revision of this test is widely employed today and will be more fully considered in Chapter 9. Of special interest, too, is the first Kuhlmann-Binet revision, which extended the scale downward to the age level of 3 months (Kuhlmann, 1912). This scale represents one of the earliest efforts to develop preschool and infant tests of intelligence.

GROUP TESTING

The Binet tests, as well as all their revisions, are *individual scales* in the sense that they can be administered to only one person at a time. Many of the tests in these scales require oral responses from the subject or necessitate the manipulation of materials. Some call for individual timing of responses. For these and other reasons, such tests are not adapted to group administration. Another characteristic of the Binet type of test is that it requires a highly trained examiner. Such tests are essentially clinical instruments, suited to the intensive study of individual cases.

Group testing, like the first Binet scale, was developed to meet a pressing practical need. When the United States entered World War I in 1917, a committee was appointed by the American Psychological Association to consider ways in which psychology might assist in the conduct of the war. This committee, under the direction of Robert M. Yerkes, recognized the need for the rapid classification of the million and a half recruits with respect to general intellectual level. Such information was relevant to many administrative decisions, including rejection or discharge from military service, assignment to different types of service, or admission to officer-training camps. It was in this setting that the first group intelligence test was developed. In this task, the Army psychologists drew on all available test materials, and especially on an unpublished group intelligence test prepared by Arthur S. Otis, which he turned over to the Army. A major contribution of Otis's test, which he designed while a student in one of Terman's graduate courses, was the introduction of multiple-choice and other "objective" item types.

The tests finally developed by the Army psychologists came to be known as the Army Alpha and the Army Beta. The former was designed for general routine testing; the latter was a nonlanguage scale employed with illiterates and with foreign-born recruits who were unable to take a test in English. Both tests were suitable for administration to large groups.

Shortly after the termination of World War I, the Army tests were released for civilian use. Not only did the Army Alpha and Army Beta themselves pass through many revisions, the latest of which are even now in use, but they also served as models for most group intelligence tests. The testing movement underwent a tremendous spurt of growth. Soon group intelligence tests were being devised for all ages and types of persons, from preschool children to graduate students. Large-scale testing programs, previously impossible, were now being launched with zestful optimism. Because group tests were designed as mass testing instruments, they not only permitted the simultaneous examination of large groups but also simplified the instructions and administration procedures so as to demand a minimum of training on the part of the examiner. Schoolteachers began to give intelligence tests to their classes. College students were routinely examined prior to admission. Extensive studies of special adult groups, such as prisoners, were undertaken. And soon the general public became IQ-conscious.

The application of such group intelligence tests far outran their technical improvement. That the tests were still crude instruments was often forgotten in the rush of gathering scores and drawing practical conclusions from the results. When the tests failed to meet unwarranted expectations, skepticism and hostility toward all testing often resulted. Thus, the testing boom of the twenties, based on the indiscriminate use of tests, may have done as much to retard as to advance the progress of psychological testing.

APTITUDE TESTING

Although intelligence tests were originally designed to sample a wide variety of functions in order to estimate the individual's general intellectual level, it soon became apparent that such tests were quite limited in their coverage. Not all important functions were represented. In fact, most intelligence tests were primarily measures of verbal ability and, to a lesser extent, of the ability to handle numerical and other abstract and symbolic relations. Gradually psychologists came to recognize that the term "intelligence test" was a misnomer, since only certain aspects of intelligence were measured by such tests.

To be sure, the tests covered abilities that are of prime importance in our culture. But it was realized that more precise designations, in terms of the type of information these tests are able to yield, would be preferable. For example, a number of tests that would probably have been called intelligence tests during the twenties later came to be known as scholastic aptitude tests. This shift in terminology was made in recognition of the fact that many so-called intelligence tests measure that combination of abilities demanded by academic work.

Even prior to World War I, psychologists had begun to recognize the need for tests of special aptitudes to supplement the global intelligence tests. These *special aptitude tests* were developed particularly for use in vocational counseling and in the selection and classification of industrial and military personnel. Among the most widely used are tests of mechanical, clerical, musical, and artistic aptitudes.

The critical evaluation of intelligence tests that followed their widespread and indiscriminate use during the twenties also revealed another noteworthy fact: an individual's performance on different parts of such a test often showed marked variation. This was especially apparent on group tests, in which the items are commonly segregated into subtests of relatively homogeneous content. For example, a person might score relatively high on a verbal subtest and low on a numerical subtest, or vice versa. To some extent, such internal variability is also discernible on a test like the Stanford-Binet, in which, for example, all items involving words might prove difficult for a particular individual, whereas items employing pictures or geometric diagrams may place him at an advantage.

Test users, and especially clinicians, frequently utilized such intercomparisons in order to obtain more insight into the individual's psychological make-up. Thus, not only the IQ or other global score but also scores on subtests would be examined in the evaluation of the individual case. Such a practice is not to be generally recommended, however, because intelligence tests were not designed for the purpose of differential aptitude analysis. Often the subtests being compared contain too few items to yield a stable or reliable estimate of a specific ability. As a result, the obtained difference between subtest scores might be reversed if the individual were retested on a different day or with another form of the same test. If such intraindividual comparisons are to be made, tests are needed that are specially designed to reveal differences in performance in various functions.

While the practical application of tests demonstrated the need for differential aptitude tests, a parallel development in the study of trait organization was gradually providing the means for constructing such tests. Statistical studies on the nature of intelligence had been exploring the interrelations among scores obtained by many persons on a wide variety of different tests. Such investigations were begun by the English psychologist Charles Spearman (1904, 1927) during the first decade of the present century. Subsequent methodological developments, based on the work of such American psychologists as T. L. Kelley (1928) and L. L. Thurstone (1935, 1947), as well as on that of other American and English investigators, have come to be known as "factor analysis."

The contributions that the methods of factor analysis have made to test construction will be more fully examined and illustrated in Chapter 13. For the present, it will suffice to note that the data gathered by such procedures have indicated the presence of a number of relatively independent factors, or traits. Some of these traits were represented, in varying proportions, in the traditional intelligence tests. Verbal comprehension and numerical reasoning are examples of this type of trait. Others, such as spatial, perceptual, and mechanical aptitudes, were found more often in special aptitude tests than in intelligence tests.

One of the chief practical outcomes of factor analysis was the development of *multiple aptitude batteries.* These batteries are designed to provide a measure of the individual's standing in each of a number of traits. In place of a total score or IQ, a separate score is obtained for such traits as verbal comprehension, numerical aptitude, spatial visualization, arithmetic reasoning, and perceptual speed. Such batteries thus provide a suitable instrument for making the kind of intraindividual analysis, or differential diagnosis, that clinicians had been trying for many years to obtain, with crude and often erroneous results, from intelligence tests. These batteries also incorporate into a comprehensive and systematic testing program much of the information formerly obtained from special aptitude tests, since the multiple aptitude batteries cover some of the traits not ordinarily included in intelligence tests.

Multiple aptitude batteries represent a relatively late development in the testing field. Nearly all have appeared since 1945. In this connection, the work of the military psychologists during World War II should also be noted. Much of the test research conducted in the armed services was based on factor analysis and was directed toward the construction of multiple aptitude batteries. In the Air Force, for example, special batteries were constructed for pilots, bombardiers, radio operators, range finders, and scores of other military specialists. A report of the batteries prepared in the Air Force alone occupies at least nine of the nineteen volumes devoted to the aviation psychology program during World War II (Army Air Forces, 1947–1948). Research along these lines is still in progress under the sponsorship of various branches of the armed services. A number of multiple aptitude batteries have likewise been developed for civilian use and are being widely applied in educational and vocational counseling and in personnel selection and classification. Examples of such batteries will be discussed in Chapter 13.

To avoid confusion, a point of terminology should be clarified. The term "aptitude test" has been traditionally employed to refer to tests measuring relatively homogeneous and clearly defined segments of ability; the term "intelligence test" customarily refers to more heterogeneous tests yielding a single global score such as an IQ. Special aptitude tests typically measure a single aptitude. Multiple aptitude batteries measure a number of aptitudes but provide a profile of scores, one for each aptitude.

STANDARDIZED ACHIEVEMENT TESTS

While psychologists were busy developing intelligence and aptitude tests, traditional school examinations were undergoing a number of technical improvements (Caldwell & Courtis, 1923; Ebel & Damrin, 1960). An important step in this direction was taken by the Boston public schools in 1845, when written examinations were substituted for the oral interrogation of students by visiting examiners. Commenting on this innovation, Horace Mann cited arguments remarkably similar to those used much later to justify the replacement of essay questions by objective multiple-choice items. The written examinations, Mann noted, put all students in a uniform situation, permitted a wider coverage of content, reduced the chance element in question choice, and eliminated the possibility of favoritism on the examiner's part.

After the turn of the century, the first standardized tests for measuring the outcomes of school instruction began to appear. Spearheaded by the work of E. L. Thorndike, these tests utilized measurement principles developed in the psychological laboratory. Examples include scales for rating the quality of handwriting and written compositions, as well as tests in spelling, arithmetic computation, and arithmetic reasoning. Still later came the achievement batteries, initiated by the publication of the first edition of the Stanford Achievement Test in 1923. Its authors were three early leaders in test development: Truman L. Kelley, Giles M. Ruch, and Lewis M. Terman. Foreshadowing many characteristics of modern testing, this battery provided comparable measures of performance in different school subjects, evaluated in terms of a single normative group.

At the same time, evidence was accumulating regarding the lack of agreement among teachers in grading essay tests. By 1930 it was widely recognized that essay tests were not only more time-consuming for examiners and examinees, but also yielded less reliable results than the "new type" of objective items. As the latter came into increasing use in standardized achievement tests, there was a growing emphasis on the design of items to test the understanding and application of knowledge and other broad educational objectives. The decade of the 1930s also witnessed the introduction of test-scoring machines, for which the new objective tests could be readily adapted.

The establishment of statewide, regional, and national testing programs was another noteworthy parallel development. Probably the best known of these programs is that of the College Entrance Examination Board (CEEB). Established at the turn of the century to reduce duplication in the examining of entering college freshmen, this program has undergone profound changes in its testing procedures and in the number and nature of participating colleges—changes that reflect intervening developments in both testing and education. In 1947, the testing functions of the CEEB were merged with those of the Carnegie Corporation and the American Council on Education to form Educational Testing Service (ETS). In subsequent years, ETS has assumed responsibility for a growing number of testing programs on behalf of universities, professional schools, government agencies, and other institutions. Mention should also be made of the American College Testing Program established in 1959 to screen applicants to colleges not included in the CEEB program, and of several national testing programs for the selection of highly talented students for scholarship awards.

Achievement tests are used not only for educational purposes but also in the selection of applicants for industrial and government jobs. Mention has already been made of the systematic use of civil service examinations in the Chinese empire, dating from 1115 B.C. In modern times, selection of government employees by examination was introduced in European countries in the late eighteenth and early nineteenth centuries. The United States Civil Service Commission installed competitive examinations as a regular procedure in 1883 (Kavruck, 1956). Test construction techniques developed during and prior to World War I were introduced into the examination program of the United States Civil Service with the appointment of L. J. O'Rourke as director of the newly established research division in 1922.

As more and more psychologists trained in psychometrics participated in the construction of standardized achievement tests, the technical aspects of achievement tests increasingly came to resemble those of intelligence and aptitude tests. Procedures for constructing and evaluating all these tests have much in common. The increasing efforts to prepare achievement tests that would measure the attainment of broad educational goals, as contrasted to the recall of factual minutiae, also made the content of achievement tests resemble more closely that of intelligence tests. Today the difference between these two types of tests is chiefly one of degree of specificity of content and extent to which the test presupposes a designated course of prior instruction.

MEASUREMENT OF PERSONALITY

Another area of psychological testing is concerned with the affective or nonintellectual aspects of behavior. Tests designed for this purpose are commonly known as personality tests, although some psychologists prefer to use the term personality in a broader sense, to refer to the entire individual. Intellectual as well as nonintellectual traits would thus be included under this heading. In the terminology of psychological testing, however, the designation "personality test" most often refers to measures of such characteristics as emotional adjustment, interpersonal relations, motivation, interests, and attitudes.

An early precursor of personality testing may be recognized in Kraepelin's use of the free association test with abnormal patients. In this test, the subject is given specially selected stimulus words and is required to respond to each with the first word that comes to mind. Kraepelin (1892) also employed this technique to study the psychological effects of fatigue, hunger, and drugs and concluded that all these agents increase the relative frequency of superficial associations. Sommer (1894), also writing during the last decade of the nineteenth century, suggested that the free association test might be used to differentiate between the various forms of mental disorder. The free association technique has subsequently been utilized for a variety of testing purposes and is still currently employed. Mention should also be made of the work of Galton, Pearson, and Cattell in the development of standardized questionnaire and rating-scale techniques. Although originally devised for other purposes, these procedures were eventually employed by others in constructing some of the most common types of current personality tests.

The prototype of the personality questionnaire, or *self-report inventory*, is the Personal Data Sheet developed by Woodworth during World War I (DuBois, 1970; Symonds, 1931, Ch. 5; Goldberg, 1971). This test was designed as a rough screening device for identifying seriously neurotic men who would be unfit for military service. The inventory consisted of a number of questions dealing with common neurotic symptoms, which the individual answered about himself. A total score was obtained by counting the number of symptoms reported. The Personal Data Sheet was not completed early enough to permit its operational use before the war ended. Immediately after the war, however, civilian forms were prepared, including a special form for use with children. The Woodworth Personal Data Sheet, moreover, served as a model for most subsequent emotional adjustment inventories. In some of these questionnaires, an attempt was made to subdivide emotional adjustment into more specific forms, such as home adjustment, school adjustment, and vocational adjustment. Other tests concentrated more intensively on a narrower area of behavior or were concerned with more distinctly social responses, such as dominance-submission in interpersonal contacts. A later development was the construction of tests for quantifying the expression of interests and attitudes. These tests, too, were based essentially on questionnaire techniques.

Another approach to the measurement of personality is through the application of *performance* or *situational tests*. In such tests, the subject has a task to perform whose purpose is often disguised. Most of these tests simulate everyday-life situations quite closely. The first extensive application of such techniques is to be found in the tests developed in the late twenties and early thirties by Hartshorne, May, and their associates (1928, 1929, 1930). This series, standardized on schoolchildren, was concerned with such behavior as cheating, lying, stealing, cooperativeness, and persistence. Objective, quantitative scores could be obtained on each of a large number of specific tests. A more recent illustration, for the adult level, is provided by the series of situational tests developed during World War II in the Assessment Program of the Office of Strategic Services (OSS, 1948). These tests were concerned with relatively complex and subtle social and emotional behavior and required rather elaborate facilities and trained personnel for their administration. The interpretation of the subject's responses, moreover, was relatively subjective.

Projective techniques represent a third approach to the study of personality and one that has shown phenomenal growth, especially among clinicians. In such tests, the subject is given a relatively unstructured task that permits wide latitude in its solution. The assumption underlying such methods is that the individual will project his characteristic modes of response into such a task. Like the performance and situational tests, projective techniques are more or less disguised in their purpose, thereby reducing the chances that the subject can deliberately create a desired impression. The previously cited free association test represents one of the earliest types of projective techniques. Sentence-completion tests have also been used in this manner. Other tasks commonly employed in projective techniques include drawing, arranging toys to create a scene, extemporaneous dramatic play, and interpreting pictures or inkblots.

All available types of personality tests present serious difficulties, both practical and theoretical. Each approach has its own special advantages and disadvantages. On the whole, personality testing has lagged far behind aptitude testing in its positive accomplishments. But such lack of progress is not to be attributed to insufficient effort. Research on the measurement of personality has attained impressive proportions since 1950, and many ingenious devices and technical improvements are under investigation. It is rather the special difficulties encountered in the measurement of personality that account for the slow advances in this area.

SOURCES OF INFORMATION ABOUT TESTS

Psychological testing is in a state of rapid change. There are shifting orientations, a constant stream of new tests, revised forms of old tests, and additional data that may refine or alter the interpretation of scores on existing tests. The accelerating rate of change, together with the vast number of available tests, makes it impracticable to survey specific tests in any single text. More intensive coverage of testing instruments and problems in special areas can be found in books dealing with the use of tests in such fields as counseling, clinical practice, personnel selection, and education. References to such publications are given in the appropriate chapters of this book. In order to keep abreast of current developments, however, anyone working with tests needs to be familiar with more direct sources of contemporary information about tests.

One of the most important sources is the series of *Mental Measurements Yearbooks* (MMY) edited by Buros (1972). These yearbooks cover nearly all commercially available psychological, educational, and vocational tests published in English. The coverage is especially complete for paper-and-pencil tests. Each yearbook includes tests published during a specified period, thus supplementing rather than supplanting the earlier yearbooks. The *Seventh Mental Measurements Yearbook*, for example, is concerned principally with tests appearing between 1964 and 1970. Tests of continuing interest, however, may be reviewed repeatedly in successive yearbooks, as new data accumulate from pertinent research. The earliest publications in this series were merely bibliographies of tests. Beginning in 1938, however, the yearbook assumed its current form, which includes critical reviews of most of the tests by one or more test experts, as well as a complete list of published references pertaining to each test. Routine information regarding publisher, price, forms, and age of subjects for whom the test is suitable is also regularly given.

A comprehensive bibliography covering all types of published tests available in English-speaking countries is provided by *Tests in Print* (Buros, 1974). Two related sources are *Reading Tests and Reviews* (Buros, 1968) and *Personality Tests and Reviews* (Buros, 1970). Both include a number of tests not found in any volume of the MMY, as well as master indexes that facilitate the location of tests in the MMY. Reviews of specific tests are also published in several psychological and educational journals, such as the *Journal of Educational Measurement* and the *Journal of Counseling Psychology*.

Since 1970, several sourcebooks have appeared which provide information about unpublished or little known instruments, largely supplementing the material listed in the MMY. A comprehensive survey of such instruments can be found in *A Sourcebook for Mental Health Measures* (Comrey, Backer, & Glaser, 1973). Containing approximately 1,100 abstracts, this sourcebook includes tests, questionnaires, rating scales, and other devices for assessing both aptitude and personality variables in adults and children. Another similar reference is entitled *Measures for Psychological Assessment* (Chun, Cobb, & French, 1975). For each of 3,000 measures, this volume gives the original source as well as an annotated bibliography of the studies in which the measure was subsequently used. The entries were located through a search of 26 measurement-related journals for the years 1960 to 1970.

Information on assessment devices suitable for children from birth to 12 years is summarized in *Tests and Measurements in Child Development: A Handbook* (Johnson & Bommarito, 1971). Covering only tests not listed in the MMY, this handbook describes instruments located through an intensive journal search spanning a ten-year period. Selection criteria included availability of the test to professionals, adequate instructions for administration and scoring, sufficient length, and convenience of use (i.e., not requiring expensive or elaborate equipment). A still more specialized collection covers measures of social and emotional development applicable to children between the ages of 3 and 6 years (Walker, 1973).

Finally, it should be noted that the most direct source of information regarding specific current tests is provided by the catalogues of test publishers and by the manual that accompanies each test. A comprehensive list of test publishers, with addresses, can be found in the latest *Mental Measurements Yearbook.* For ready reference, the names and addresses of some of the larger American publishers and distributors of psychological tests are given in Appendix D. Catalogues of current tests can be obtained from each of these publishers on request. Manuals and specimen sets of tests can be purchased by qualified users.

The test manual should provide the essential information required for administering, scoring, and evaluating a particular test. In it should be found full and detailed instructions, scoring key, norms, and data on reliability and validity. Moreover, the manual should report the number and nature of subjects on whom norms, reliability, and validity were established, the methods employed in computing indices of reliability and validity, and the specific criteria against which validity was checked. In the event that the necessary information is too lengthy to fit conveniently into the manual, references to the printed sources in which such information can be readily located should be given. The manual should, in other words, enable the test user to evaluate the test before choosing it for his specific purpose. It might be added that many test manuals still fall short of this goal. But some of the larger and more professionally oriented test publishers are giving increasing attention to the preparation of manuals that meet adequate scientific standards. An enlightened public of test users provides the firmest assurance that such standards will be maintained and improved in the future.

A succinct but comprehensive guide for the evaluation of psychological tests is to be found in *Standards for Educational and Psychological Tests* (1974), published by the American Psychological Association. These standards represent a summary of recommended practices in test construction, based on the current state of knowledge in the field. They are concerned with the information about validity, reliability, norms, and other test characteristics that ought to be reported in the manual. In their latest revision, the *Standards* also provide a guide for the proper use of tests and for the correct interpretation and application of test results. Relevant portions of the *Standards* will be cited in the following chapters, in connection with the appropriate topics.

CHAPTER 2

Nature and Use of Psychological Tests

THE HISTORICAL introduction in Chapter 1 has already suggested some of the many uses of psychological tests, as well as the wide diversity of available tests. Although the general public may still associate psychological tests most closely with "IQ tests" and with tests designed to detect emotional disorders, these tests represent only a small proportion of the available types of instruments. The major categories of psychological tests will be discussed and illustrated in Parts 3, 4, and 5, which cover tests of general intellectual level, traditionally called intelligence tests; tests of separate abilities, including multiple aptitude batteries, tests of special aptitudes, and achievement tests; and personality tests, concerned with measures of emotional and motivational traits, interpersonal behavior, interests, attitudes, and other noncognitive characteristics.

In the face of such diversity in nature and purpose, what are the common differentiating characteristics of psychological tests? How do psychological tests differ from other methods of gathering information about individuals? The answer is to be found in certain fundamental features of both the construction and use of tests. It is with these features that the present chapter is concerned.

WHAT IS A PSYCHOLOGICAL TEST?

BEHAVIOR SAMPLE. A psychological test is essentially an objective and standardized measure of a sample of behavior. Psychological tests are like tests in any other science, insofar as observations are made on a small but carefully chosen *sample* of an individual's behavior. In this respect, the psychologist proceeds in much the same way as the chemist who tests a patient's blood or a community's water supply by analyzing one or more samples of it. If the psychologist wishes to test the extent of a child's vocabulary, a clerk's ability to perform arithmetic computations, or a pilot's eye-hand coordination, he examines their performance with a representative set of words, arithmetic problems, or motor tests. Whether or not the test adequately covers the behavior under consideration obviously depends on the number and nature of items in the sample. For example, an arithmetic test consisting of only five problems, or one including only multiplication items, would be a poor measure of the individual's computational skill. A vocabulary test composed entirely of baseball terms would hardly provide a dependable estimate of a child's total range of vocabulary.

The *diagnostic* or *predictive value* of a psychological test depends on the degree to which it serves as an indicator of a relatively broad and significant area of behavior. Measurement of the behavior sample directly covered by the test is rarely, if ever, the goal of psychological testing. The child's knowledge of a particular list of 50 words is not, in itself, of great interest. Nor is the job applicant's performance on a specific set of 20 arithmetic problems of much importance. If, however, it can be demonstrated that there is a close correspondence between the child's knowledge of the word list and his total mastery of vocabulary, or between the applicant's score on the arithmetic problems and his computational performance on the job, then the tests are serving their purpose.

It should be noted in this connection that the test items need not resemble closely the behavior the test is to predict. It is only necessary that an empirical correspondence be demonstrated between the two. The degree of similarity between the test sample and the predicted behavior may vary widely. At one extreme, the test may coincide completely with a part of the behavior to be predicted. An example might be a foreign vocabulary test in which the students are examined on 20 of the 50 new words they have studied; another example is provided by the road test taken prior to obtaining a driver's license. A lesser degree of similarity is illustrated by many vocational aptitude tests administered prior to job training, in which there is only a moderate resemblance between the tasks performed on the job and those incorporated in the test. At the other extreme one finds projective personality tests such as the Rorschach inkblot test, in which an attempt is made to predict from the subject's associations to inkblots how he will react to other people, to emotionally toned stimuli, and to other complex, everyday-life situations. Despite their superficial differences, all these tests consist of samples of the individual's behavior. And each must prove its worth by an empirically demonstrated correspondence between the subject's performance on the test and in other situations.

Whether the term "diagnosis" or the term "prediction" is employed in this connection also represents a minor distinction. Prediction commonly connotes a temporal estimate, the individual's future performance on a job, for example, being forecast from his present test performance. In a broader sense, however, even the diagnosis of present condition, such as mental retardation or emotional disorder, implies a prediction of what the individual will do in situations other than the present test. It is logically simpler to consider all tests as behavior samples from which predictions regarding other behavior can be made. Different types of tests can then be characterized as variants of this basic pattern.

Another point that should be considered at the outset pertains to the concept of *capacity*. It is entirely possible, for example, to devise a test for predicting how well an individual can learn French before he has even begun the study of French. Such a test would involve a sample of the types of behavior required to learn the new language, but would in itself presuppose no knowledge of French. It could then be said that this test measures the individual's "capacity" or "potentiality" for learning French. Such terms should, however, be used with caution in reference to psychological tests. Only in the sense that a present behavior sample can be used as an indicator of other, future behavior can we speak of a test measuring "capacity." No psychological test can do more than measure behavior. Whether such behavior can serve as an effective index of other behavior can be determined only by empirical try-out.

STANDARDIZATION. It will be recalled that in the initial definition a psychological test was described as a standardized measure. Standardization implies *uniformity of procedure* in administering and scoring the test. If the scores obtained by different individuals are to be comparable, testing conditions must obviously be the same for all. Such a requirement is only a special application of the need for controlled conditions in all scientific observations. In a test situation, the single independent variable is usually the individual being tested.

In order to secure uniformity of testing conditions, the test constructor provides detailed directions for administering each newly developed test. The formulation of such directions is a major part of the standardization of a new test. Such standardization extends to the exact materials employed, time limits, oral instructions to subjects, preliminary demonstrations, ways of handling queries from subjects, and every other detail of the testing situation. Many other, more subtle factors may influence the subject's performance on certain tests. Thus, in giving instructions or presenting problems orally, consideration must be given to the rate of speaking, tone of voice, inflection, pauses, and facial expression. In a test involving the detection of absurdities, for example, the correct answer may be given away by smiling or pausing when the crucial word is read. Standardized testing procedure, from the examiner's point of view, will be discussed further in a later section of this chapter dealing with problems of test administration.

Another important step in the standardization of a test is the establishment of *norms*. Psychological tests have no predetermined standards of passing or failing; an individual's score is evaluated by comparing it with the scores obtained by others. As its name implies, a norm is the normal or average performance. Thus, if normal 8-year-old children complete 12 out of 50 problems correctly on a particular arithmetic reasoning test, then the 8-year-old norm on this test corresponds to a score of 12. The latter is known as the raw score on the test. It may be expressed as number of correct items, time required to complete a task, number of errors, or some other objective measure appropriate to the content of the test. Such a raw score is meaningless until evaluated in terms of a suitable set of norms.

In the process of standardizing a test, it is administered to a large, representative sample of the type of subjects for whom it is designed. This group, known as the standardization sample, serves to establish the norms. Such norms indicate not only the average performance but also the relative frequency of varying degrees of deviation above and below the average. It is thus possible to evaluate different degrees of superiority and inferiority. The specific ways in which such norms may be expressed will be considered in Chapter 4. All permit the designation of the individual's position with reference to the normative or standardization sample.

It might also be noted that norms are established for personality tests in essentially the same way as for aptitude tests. The norm on a personality test is not necessarily the most desirable or "ideal" performance, any more than a perfect or errorless score is the norm on an aptitude test. On both types of tests, the norm corresponds to the performance of typical or average individuals. On dominance-submission tests, for example, the norm falls at an intermediate point representing the degree of dominance or submission manifested by the average individual. Similarly, in an emotional adjustment inventory, the norm does not ordinarily correspond to a complete absence of unfavorable or maladaptive responses, since a few such responses occur in the majority of "normal" individuals in the standardization sample. It is thus apparent that psychological tests, of whatever type, are based on empirically established norms.

OBJECTIVE MEASUREMENT OF DIFFICULTY. Reference to the definition of a psychological test with which this discussion opened will show that such a test was characterized as an objective as well as a standardized measure. In what specific ways are such tests objective? Some aspects of the objectivity of psychological tests have already been touched on in the discussion of standardization. Thus, the administration, scoring, and interpretation of scores are objective insofar as they are independent of the subjective judgment of the individual examiner. Any one individual should theoretically obtain the identical score on a test regardless of who happens to be his examiner. This is not entirely so, of course, since perfect standardization and objectivity have not been attained in practice. But at least such objectivity is the goal of test construction and has been achieved to a reasonably high degree in most tests.

There are other major ways in which psychological tests can be properly described as objective. The determination of the difficulty level of an item or of a whole test is based on objective, empirical procedures. When Binet and Simon prepared their original, 1905 scale for the measurement of intelligence, they arranged the 30 items of the scale in order of increasing difficulty. Such difficulty, it will be recalled, was determined by trying out the items on 50 normal and a few mentally retarded children. The items correctly solved by the largest number of children were, *ipso facto,* taken to be the easiest; those passed by relatively few children were regarded as more difficult items. By this procedure, an empirical order of difficulty was established. This early example typifies the objective measurement of difficulty level, which is now common practice in psychological test construction.

Not only the arrangement but also the selection of items for inclusion in a test can be determined by the proportion of subjects in the trial samples who pass each item. Thus, if there is a bunching of items at the easy or difficult end of the scale, some items can be discarded. Similarly, if items are sparse in certain portions of the difficulty range, new items can be added to fill the gaps. More technical aspects of item analysis will be considered in Chapter 8.

RELIABILITY. How good is this test? Does it really work? These questions could—and occasionally do—result in long hours of futile discussion. Subjective opinions, hunches, and personal biases may lead, on the one hand, to extravagant claims regarding what a particular test can accomplish and, on the other hand, to stubborn rejection. The only way questions such as these can be conclusively answered is by empirical trial. The *objective evaluation* of psychological tests involves primarily the determination of the reliability and the validity of the test in specified situations.

As used in psychometrics, the term reliability always means consistency. Test reliability is the consistency of scores obtained by the same persons when retested with the identical test or with an equivalent form of the test. If a child receives an IQ of 110 on Monday and an IQ of 80 when retested on Friday, it is obvious that little or no confidence can be put in either score. Similarly, if in one set of 50 words an individual identifies 40 correctly, whereas in another, supposedly equivalent set he gets a score of only 20 right, then neither score can be taken as a dependable index of his verbal comprehension. To be sure, in both illustrations it is possible that only one of the two scores is in error, but this could be demonstrated only by further retests. From the given data, we can conclude only that both scores cannot be right. Whether one or neither is an adequate estimate of the individual's ability in vocabulary cannot be established without additional information.

Before a psychological test is released for general use, a thorough, objective check of its reliability should be carried out. The different types of test reliability, as well as methods of measuring each, will be considered in Chapter 5. Reliability can be checked with reference to temporal fluctuations, the particular selection of items or behavior sample constituting the test, the role of different examiners or scorers, and other aspects of the testing situation. It is essential to specify the type of reliability and the method employed to determine it, because the same test may vary in these different aspects. The number and nature of individuals on whom reliability was checked should likewise be reported. With such information, the test user can predict whether the test will be about equally reliable for the group with which he expects to use it, or whether it is likely to be more reliable or less reliable.

VALIDITY. Undoubtedly the most important question to be asked about any psychological test concerns its validity, i.e., the degree to which the test actually measures what it purports to measure. Validity provides a direct check on how well the test fulfills its function. The determination of validity usually requires independent, external *criteria* of whatever the test is designed to measure. For example, if a medical aptitude test is to be used in selecting promising applicants for medical school, ultimate success in medical school would be a criterion. In the process of validating such a test, it would be administered to a large group of students at the time of their admission to medical school. Some measure of performance in medical school would eventually be obtained for each student on the basis of grades, ratings by instructors, success or failure in completing training, and the like. Such a composite measure constitutes the criterion with which each student's initial test score is to be correlated. A high correlation, or *validity coefficient*, would signify that those individuals who scored high on the test had been relatively successful in medical school, whereas those scoring low on the test had done poorly in medical school. A low correlation would indicate little correspondence between test score and criterion measure and hence poor validity for the test. The validity coefficient enables us to determine how closely the criterion performance could have been predicted from the test scores.

In a similar manner, tests designed for other purposes can be validated against appropriate criteria. A vocational aptitude test, for example, can be validated against on-the-job success of a trial group of new employees. A pilot aptitude battery can be validated against achievement in flight training. Tests designed for broader and more varied uses are validated against a number of criteria and their validity can be established only by the gradual accumulation of data from many different kinds of investigations.

The reader may have noticed an apparent paradox in the concept of test validity. If it is necessary to follow up the subjects or in other ways to obtain independent measures of what the test is trying to predict, why not dispense with the test? The answer to this riddle is to be found in the distinction between the validation group on the one hand and the groups on which the test will eventually be employed for operational purposes on the other. Before the test is ready for use, its validity must be established on a representative sample of subjects. The scores of these persons are not themselves employed for operational purposes but serve only in the process of testing the test. If the test proves valid by this method, it can then be used on other samples in the absence of criterion measures.

It might still be argued that we would need only to wait for the criterion measure to mature, to become available, on *any* group in order to obtain the information that the test is trying to predict. But such a procedure would be so wasteful of time and energy as to be prohibitive in most instances. Thus, we could determine which applicants will succeed on a job or which students will satisfactorily complete college by admitting all who apply and waiting for subsequent developments! It is the very wastefulness of this procedure—and its deleterious emotional impact on individuals—that tests are designed to minimize. By means of tests, the person's present level of prerequisite skills, knowledge, and other relevant characteristics can be assessed with a determinable margin of error. The more valid and reliable the test, the smaller will be this margin of error.

The special problems encountered in determining the validity of different types of tests, as well as the specific criteria and statistical procedures employed, will be discussed in Chapters 6 and 7. One further point, however, should be considered at this time. Validity tells us more than the degree to which the test is fulfilling its function. It actually tells us *what* the test is measuring. By studying the validation data, we can objectively determine what the test is measuring. It would thus be more accurate to define validity as the extent to which we know what the test measures. The interpretation of test scores would undoubtedly be clearer and less ambiguous if tests were regularly named in terms of the criterion data through which they had been validated. A tendency in this direction can be recognized in such test labels as "scholastic aptitude test" and "personnel classification test" in place of the vague title "intelligence test."

REASONS FOR CONTROLLING THE USE OF PSYCHOLOGICAL TESTS

"May I have a Stanford-Binet blank? My nephew has to take it next week for admission to School X and I'd like to give him some practice so he can pass."

"To improve the reading program in our school, we need a culture-free IQ test that measures each child's innate potential."

"Last night I answered the questions in an intelligence test published in a magazine and I got an IQ of 80—I think psychological tests are silly."

"My roommate is studying psych. She gave me a personality test and I came out neurotic. I've been too upset to go to class ever since."

"Last year you gave a new personality test to our employees for research purposes. We would now like to have the scores for their personnel folders."

The above remarks are not imaginary. Each is based on a real incident, and the list could easily be extended by any psychologist. Such remarks illustrate potential misuses or misinterpretations of psychological tests in such ways as to render the tests worthless or to hurt the individual. Like any scientific instrument or precision tool, psychological tests must be properly used to be effective. In the hands of either the unscrupulous or the well-meaning but uninformed user, such tests can cause serious damage.

There are two principal reasons for controlling the use of psychological tests: (*a*) to prevent general familiarity with test content, which would invalidate the test and (*b*) to ensure that the test is used by a qualified examiner. Obviously, if an individual were to memorize the correct responses on a test of color blindness, such a test would no longer be a measure of color vision for him. Under these conditions, the test would be completely invalidated. Test content clearly has to be restricted in order to forestall deliberate efforts to fake scores.

In other cases, however, the effect of familiarity may be less obvious, or the test may be invalidated in good faith by misinformed persons. A schoolteacher, for example, may give her class special practice in problems closely resembling those on an intelligence test, "so that the pupils will be well prepared to take the test." Such an attitude is simply a carryover from the usual procedure of preparing for a school examination.

When applied to an intelligence test, however, it is likely that such specific training or coaching will raise the scores on the test without appreciably affecting the broader area of behavior the test tries to sample. Under such conditions, the validity of the test as a predictive instrument is reduced.

The need for a qualified examiner is evident in each of the three major aspects of the testing situation—selection of the test, administration and scoring, and interpretation of scores. Tests cannot be chosen like lawn mowers, from a mail-order catalogue. They cannot be evaluated by name, author, or other easy marks of identification. To be sure, it requires no psychological training to consider such factors as cost, bulkiness and ease of transporting test materials, testing time required, and ease and rapidity of scoring. Information on these practical points can usually be obtained from a test catalogue and should be taken into account in planning a testing program. For the test to serve its function, however, an evaluation of its technical merits in terms of such characteristics as validity, reliability, difficulty level, and norms is essential. Only in such a way can the test user determine the appropriateness of any test for his particular purpose and its suitability for the type of persons with whom he plans to use it.

The introductory discussion of test standardization earlier in this chapter has already suggested the importance of a trained examiner. An adequate realization of the need to follow instructions precisely, as well as a thorough familiarity with the standard instructions, is required if the test scores obtained by different examiners are to be comparable or if any one individual's score is to be evaluated in terms of the published norms. Careful control of testing conditions is also essential. Similarly, incorrect or inaccurate scoring may render the test score worthless. In the absence of proper checking procedures, scoring errors are far more likely to occur than is generally realized.

The proper interpretation of test scores requires a thorough understanding of the test, the individual, and the testing conditions. What is being measured can be objectively determined only by reference to the specific procedures in terms of which the particular test was validated. Other information, pertaining to reliability, nature of the group on which norms were established, and the like, is likewise relevant. Some background data regarding the individual being tested are essential in interpreting any test score. The same score may be obtained by different persons for very different reasons. The conclusions to be drawn from such scores would therefore be quite dissimilar. Finally, some consideration must also be given to special factors that may have influenced a particular score, such as unusual testing conditions, temporary emotional or physical state of the subject, and extent of the subject's previous experience with tests.

TEST ADMINISTRATION

The basic rationale of testing involves generalization from the behavior sample observed in the testing situation to behavior manifested in other, nontest situations. A test score should help us to predict how the client will feel and act outside the clinic, how the student will achieve in college courses, and how the applicant will perform on the job. Any influences that are specific to the test situation constitute error variance and reduce test validity. It is therefore important to identify any test-related influences that may limit or impair the generalizability of test results.

A whole volume could easily be devoted to a discussion of desirable procedures of test administration. But such a survey falls outside the scope of the present book. Moreover, it is more practicable to acquire such techniques within specific settings, because no one person would normally be concerned with all forms of testing, from the examination of infants to the clinical testing of psychotic patients or the administration of a mass testing program for military personnel. The present discussion will therefore deal principally with the common rationale of test administration rather than with specific questions of implementation. For detailed suggestions regarding testing procedure, see Palmer (1970), Sattler (1974), and Terman and Merrill (1960) for individual testing, and Clemans (1971) for group testing.

ADVANCE PREPARATION OF EXAMINERS. The most important requirement for good testing procedure is advance preparation. In testing there can be no emergencies. Special efforts must therefore be made to foresee and forestall emergencies. Only in this way can uniformity of procedure be assured.

Advance preparation for the testing session takes many forms. Memorizing the exact verbal instructions is essential in most individual testing. Even in a group test in which the instructions are read to the subjects, some previous familiarity with the statements to be read prevents misreading and hesitation and permits a more natural, informal manner during test administration. The preparation of test materials is another important preliminary step. In individual testing and especially in the administration of performance tests, such preparation involves the actual layout of the necessary materials to facilitate subsequent use with a minimum of search or fumbling. Materials should generally be placed on a table near the testing table so that they are within easy reach of the examiner but do not distract the subject. When apparatus is employed, frequent periodic checking and calibration may be necessary. In group testing, all test blanks, answer sheets, special pencils, or other materials needed should be carefully counted, checked, and arranged in advance of the testing day.

Thorough familiarity with the specific testing procedure is another important prerequisite in both individual and group testing. For individual testing, supervised training in the administration of the particular test is usually essential. Depending upon the nature of the test and the type of subjects to be examined, such training may require from a few demonstration and practice sessions to over a year of instruction. For group testing, and especially in large-scale projects, such preparation may include advance briefing of examiners and proctors, so that each is fully informed about the functions he is to perform. In general, the examiner reads the instructions, takes care of timing, and is in charge of the group in any one testing room. The proctors hand out and collect test materials, make certain that subjects are following instructions, answer individual questions of subjects within the limitations specified in the manual, and prevent cheating.

TESTING CONDITIONS. Standardized procedure applies not only to verbal instructions, timing, materials, and other aspects of the tests themselves but also to the testing environment. Some attention should be given to the selection of a suitable testing room. This room should be free from undue noise and distraction and should provide adequate lighting, ventilation, seating facilities, and working space for the subjects. Special steps should also be taken to prevent interruptions during the test. Posting a sign on the door to indicate that testing is in progress is effective, provided all personnel have learned that such a sign means no admittance under any circumstances. In the testing of large groups, locking the doors or posting an assistant outside each door may be necessary to prevent the entrance of late-comers.

It is important to realize the extent to which testing conditions may influence scores. Even apparently minor aspects of the testing situation may appreciably alter performance. Such a factor as the use of desks or of chairs with desk arms, for example, proved to be significant in a group testing project with high school students, the groups using desks tending to obtain higher scores (Kelley, 1943; Traxler & Hilkert, 1942). There is also evidence to show that the type of answer sheet employed may affect test scores (Bell, Hoff, & Hoyt, 1964). Because of the establishment of independent test-scoring and data-processing agencies that provide their own machine-scorable answer sheets, examiners sometimes administer group tests with answer sheets other than those used in the standardization sample. In the absence of empirical verification, the equivalence of these answer sheets cannot be assumed. The Differential Aptitude Tests, for example, may be administered with any of five different answer sheets. On the Clerical Speed and Accuracy Test of this battery, separate norms are provided for three of the five answer sheets, because they were found to yield substantially different scores than those obtained with the answer sheets used by the standardization sample.

In testing children below the fifth grade, the use of *any* separate answer sheet may significantly lower test scores (Metropolitan Achievement Test Special Report, 1973). At these grade levels, having the child mark the answers in the test booklet itself is generally preferable.

Many other, more subtle testing conditions have been shown to affect performance on ability as well as personality tests. Whether the examiner is a stranger or someone familiar to the subjects may make a significant difference in test scores (Sacks, 1952; Tsudzuki, Hata, & Kuze, 1957). In another study, the general manner and behavior of the examiner, as illustrated by smiling, nodding, and making such comments as "good" or "fine," were shown to have a decided effect on test results (Wickes, 1956). In a projective test requiring the subject to write stories to fit given pictures, the presence of the examiner in the room tended to inhibit the inclusion of strongly emotional content in the stories (Bernstein, 1956). In the administration of a typing test, job applicants typed at a significantly faster rate when tested alone than when tested in groups of two or more (Kirchner, 1966).

Examples could readily be multiplied. The implications are threefold. First, follow standardized procedures to the minutest detail. It is the responsibility of the test author and publisher to describe such procedures fully and clearly in the test manual. Second, record any unusual testing conditions, however minor. Third, take testing conditions into account when interpreting test results. In the intensive assessment of a person through individual testing, an experienced examiner may occasionally depart from the standardized test procedure in order to elicit additional information for special reasons. When he does so, he can no longer interpret the subject's responses in terms of the test norms. Under these circumstances, the test stimuli are used only for qualitative exploration; and the responses should be treated in the same way as any other informal behavioral observations or interview data.

RAPPORT

In psychometrics, the term "rapport" refers to the examiner's efforts to arouse the subject's interest in the test, elicit his cooperation, and ensure that he follows the standard test instructions. In ability tests, the instructions call for careful concentration on the given tasks and for putting forth one's best efforts to perform well; in personality inventories, they call for frank and honest responses to questions about one's usual behavior; in certain projective tests, they call for full reporting of associations evoked by the stimuli, without any censoring or editing of content. Still other kinds of tests may require other approaches. But in all instances, the examiner endeavors to motivate the subject to follow the instructions as fully and conscientiously as he can.

The training of examiners covers techniques for the establishment of rapport as well as those more directly related to test administration. In establishing rapport, as in other testing procedures, uniformity of conditions is essential for comparability of results. If a child is given a coveted prize whenever he solves a test problem correctly, his performance cannot be directly compared with the norms or with that of other children who are motivated only with the standard verbal encouragement or praise. Any deviation from standard motivating conditions for a particular test should be noted and taken into account in interpreting performance.

Although rapport can be more fully established in individual testing, steps can also be taken in group testing to motivate the subjects and relieve their anxiety. Specific techniques for establishing rapport vary with the nature of the test and with the age and other characteristics of the subjects. In testing preschool children, special factors to be considered include shyness with strangers, distractibility, and negativism. A friendly, cheerful, and relaxed manner on the part of the examiner helps to reassure the child. The shy, timid child needs more preliminary time to become familiar with his surroundings. For this reason it is better for the examiner not to be too demonstrative at the outset, but rather to wait until the child is ready to make the first contact. Test periods should be brief, and the tasks should be varied and intrinsically interesting to the child. The testing should be presented to the child as a game and his curiosity aroused before each new task is introduced. A certain flexibility of procedure is necessary at this age level because of possible refusals, loss of interest, and other manifestations of negativism.

Children in the first two or three grades of elementary school present many of the same testing problems as the preschool child. The game approach is still the most effective way of arousing their interest in the test. The older schoolchild can usually be motivated through an appeal to his competitive spirit and his desire to do well on tests. When testing children from educationally disadvantaged backgrounds or from different cultures, however, the examiner cannot assume they will be motivated to excel on academic tasks to the same extent as children in the standardization sample. This problem and others pertaining to the testing of persons with dissimilar experiential backgrounds will be considered further in Chapters 3, 7, and 12.

Special motivational problems may be encountered in testing emotionally disturbed persons, prisoners, or juvenile delinquents. Especially when examined in an institutional setting, such persons are likely to manifest a number of unfavorable attitudes, such as suspicion, insecurity, fear, or cynical indifference. Abnormal conditions in their past experiences are also likely to influence their test performance adversely. As a result of early failures and frustrations in school, for example, they may have developed feelings of hostility and inferiority toward academic tasks, which the tests resemble. The experienced examiner makes special efforts to establish rapport under these conditions. In any event, he must be sensitive to these special difficulties and take them into account in interpreting and explaining test performance.

In testing any school-age child or adult, one should bear in mind that every test presents an implied threat to the individual's prestige. Some reassurance should therefore be given at the outset. It is helpful to explain, for example, that no one is expected to finish or to get all the items correct. The individual might otherwise experience a mounting sense of failure as he advances to the more difficult items or finds that he is unable to finish any subtest within the time allowed.

It is also desirable to eliminate the element of surprise from the test situation as far as possible, because the unexpected and unknown are likely to produce anxiety. Many group tests provide a preliminary explanatory statement that is read to the group by the examiner. An even better procedure is to announce the tests a few days in advance and to give each subject a printed booklet that explains the purpose and nature of the tests, offers general suggestions on how to take tests, and contains a few sample items. Such explanatory booklets are regularly available to participants in large-scale testing programs such as those conducted by the College Entrance Examination Board (1974a, 1974b). The United States Employment Service has likewise developed a booklet on how to take tests, as well as a more extensive pretesting orientation technique for use with culturally disadvantaged applicants unfamiliar with tests.

More general orientation booklets are also available, such as *Meeting the Test* (Anderson, Katz, & Shimberg, 1965). A tape recording and two booklets are combined in *Test Orientation Procedure* (TOP), designed specifically for job applicants with little prior testing experience (Bennett & Doppelt, 1967). The first booklet, used together with the tape, provides general information on how to take tests; the second contains practice tests. In the absence of a tape recorder, the examiner may read the instructions from a printed script.

Adult testing presents some additional problems. Unlike the schoolchild, the adult is not so likely to work hard at a task merely because it is assigned to him. It therefore becomes more important to "sell" the purpose of the tests to the adult, although high school and college students also respond to such an appeal. Cooperation of the examinee can usually be secured by convincing him that it is in his own interests to obtain a valid score, i.e., a score correctly indicating what he can do rather than overestimating or underestimating his abilities. Most persons will understand that an incorrect decision, which might result from invalid test scores, would mean subsequent failure, loss of time, and frustration for them. This approach can serve not only to motivate the individual to try his best on ability tests but also to reduce faking and encourage frank reporting on personality inventories, because the examinee realizes that he himself would otherwise be the loser. It is certainly not in the best interests of the individual to be admitted to a course of study for which he is not qualified or assigned to a job he cannot perform or that he would find uncongenial.

TEST ANXIETY

Many of the practices designed to enhance rapport serve also to reduce test anxiety. Procedures tending to dispel surprise and strangeness from the testing situation and to reassure and encourage the subject should certainly help to lower anxiety. The examiner's own manner and a well-organized, smoothly running testing operation will contribute toward the same goal. Individual differences in test anxiety have been studied with both schoolchildren and college students (Gaudry & Spielberger, 1974; Spielberger, 1972). Much of this research was initiated by Sarason and his associates at Yale (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960). The first step was to construct a questionnaire to assess the individual's test-taking attitudes. The children's form, for example, contains items such as the following:

Do you worry a lot before taking a test?

When the teacher says she is going to find out how much you have learned, does your heart begin to beat faster?

While you are taking a test, do you usually think you are not doing well?

Of primary interest is the finding that both school achievement and intelligence test scores yielded significant negative correlations with test anxiety. Similar correlations have been found among college students (I. G. Sarason, 1961). Longitudinal studies likewise revealed an inverse relation between changes in anxiety level and changes in intelligence or achievement test performance (Hill & Sarason, 1966; Sarason, Hill, & Zimbardo, 1964).

Such findings, of course, do not indicate the direction of causal relationships. It is possible that children develop test anxiety because they perform poorly on tests and have thus experienced failure and frustration in previous test situations. In support of this interpretation is the finding that within subgroups of high scorers on intelligence tests, the negative correlation between anxiety level and test performance disappears (Denny, 1966; Feldhusen & Klausmeier, 1962). On the other hand, there is evidence suggesting that at least some of the relationship results from the deleterious effects of anxiety on test performance. In one study (Waite, Sarason, Lighthall, & Davidson, 1958), high-anxious and low-anxious children equated in intelligence test scores were given repeated trials in a learning task. Although initially equal in the learning test, the low-anxious group improved significantly more than the high-anxious.

Several investigators have compared test performance under conditions designed to evoke "anxious" and "relaxed" states. Mandler and Sarason (1952), for example, found that ego-involving instructions, such as telling subjects that everyone is expected to finish in the time allotted, had a beneficial effect on the performance of low-anxious subjects, but a deleterious effect on that of high-anxious subjects. Other studies have likewise found an interaction between testing conditions and such individual characteristics as anxiety level and achievement motivation (Lawrence, 1962; Paul & Eriksen, 1964). It thus appears likely that the relation between anxiety and test performance is nonlinear, a slight amount of anxiety being beneficial while a large amount is detrimental. Individuals who are customarily low-anxious benefit from test conditions that arouse some anxiety, while those who are customarily high-anxious perform better under more relaxed conditions.

It is undoubtedly true that a chronically high anxiety level will exert a detrimental effect on school learning and intellectual development. Such an effect, however, should be distinguished from the test-limited effects with which this discussion is concerned. To what extent does test anxiety make the individual's test performance unrepresentative of his customary performance level in nontest situations? Because of the competitive pressure experienced by college-bound high school seniors in America today, it has been argued that performance on college admission tests may be unduly affected by test anxiety. In a thorough and well-controlled investigation of this question, French (1962) compared the performance of high school students on a test given as part of the regular administration of the SAT with performance on a parallel form of the test administered at a different time under "relaxed" conditions. The instructions on the latter occasion specified that the test was given for research purposes only and scores would not be sent to any college. The results showed that performance was no poorer during the standard administration than during the relaxed administration. Moreover, the concurrent validity of the test scores against high school course grades did not differ significantly under the two conditions.

EXAMINER AND SITUATIONAL VARIABLES

Comprehensive surveys of the effects of examiner and situational variables on test scores have been prepared by S. B. Sarason (1954), Masling (1960), Moriarty (1961, 1966), Sattler and Theye (1967), Palmer (1970), and Sattler (1970, 1974). Although some effects have been demonstrated with objective group tests, most of the data have been obtained with either projective techniques or individual intelligence tests. These extraneous factors are more likely to operate with unstructured and ambiguous stimuli, as well as with difficult and novel tasks, than with clearly defined and well-learned functions. In general, children are more susceptible to examiner and situational influences than are adults; in the examination of preschool children, the role of the examiner is especially crucial. Emotionally disturbed and insecure persons of any age are also more likely to be affected by such conditions than are well-adjusted persons.

There is considerable evidence that test results may vary systematically as a function of the examiner (E. Cohen, 1965; Masling, 1960). These differences may be related to personal characteristics of the examiner, such as his age, sex, race, professional or socioeconomic status, training and experience, personality characteristics, and appearance. Several studies of these examiner variables, however, have yielded misleading or inconclusive results because the experimental designs failed to control or isolate the influence of different examiner or subject characteristics. Hence the effects of two or more variables may be confounded.

The examiner's behavior before and during test administration has also been shown to affect test results. For example, controlled investigations have yielded significant differences in intelligence test performance as a result of a "warm" versus a "cold" interpersonal relation between examiner and examinees, or a rigid and aloof versus a natural manner on the part of the examiner (Exner, 1966; Masling, 1959). Moreover, there may be significant interactions between examiner and examinee characteristics, in the sense that the same examiner characteristic or testing manner may have a different effect on different examinees as a function of the examinee's own personality characteristics. Similar interactions may occur with task variables, such as the nature of the test, the purpose of the testing, and the instructions given to the subjects. Dyer (1973) adds even more variables to this list, calling attention to the possible influence of the test givers' and the test takers' diverse perceptions of the functions and goals of testing.

Still another way in which an examiner may inadvertently affect the examinee's responses is through his own expectations. This is simply a special instance of the self-fulfilling prophecy (Rosenthal, 1966; Rosenthal & Rosnow, 1969). An experiment conducted with the Rorschach will illustrate this effect (Masling, 1965). The examiners were 14 graduate student volunteers, 7 of whom were told, among other things, that experienced examiners elicit more human than animal responses from the subjects, while the other 7 were told that experienced examiners elicit more animal than human responses. Under these conditions, the two groups of examiners obtained significantly different ratios of animal to human responses from their subjects. These differences occurred despite the fact that neither examiners nor subjects reported awareness of any influence attempt. Moreover, tape recordings of all testing sessions revealed no evidence of verbal influence on the part of any examiner. The examiners' expectations apparently operated through subtle postural and facial cues to which the subjects responded.

Apart from the examiner, other aspects of the testing situation may significantly affect test performance. Military recruits, for example, are often examined shortly after induction, during a period of intense readjustment to an unfamiliar and stressful situation. In one investigation designed to test the effect of acclimatization to such a situation on test performance, 2,724 recruits were given the Navy Classification Battery during their ninth day at the Naval Training Center (Gordon & Alf, 1960). When their scores were compared with those obtained by 2,180 recruits tested at the conventional time, during their third day, the 9-day group scored significantly higher on all subtests of the battery.

The examinees' activities immediately preceding the test may also affect their performance, especially when such activities produce emotional disturbance, fatigue, or other handicapping conditions. In an investigation with third- and fourth-grade schoolchildren, there was some evidence to suggest that IQ on the Draw-a-Man Test was influenced by the children's preceding classroom activity (McCarthy, 1944). On one occasion, the class had been engaged in writing a composition on "The Best Thing That Ever Happened to Me"; on the second occasion, they had again been writing, but this time on "The Worst Thing That Ever Happened to Me." The IQ's on the second test, following what may have been an emotionally depressing experience, averaged 4 or 5 points lower than on the first test. These findings were corroborated in a later investigation specifically designed to determine the effect of immediately preceding experience on the Draw-a-Man Test (Reichenberg-Hackett, 1953). In this study, children who had had a gratifying experience involving the successful solution of an interesting puzzle, followed by a reward of toys and candy, showed more improvement in their test scores than those who had undergone neutral or less gratifying experiences. Similar results were obtained by W. E. Davis (1969a, 1969b) with college students. Performance on an arithmetic reasoning test was significantly poorer when preceded by a failure experience on a verbal comprehension test than it was in a control group given no preceding test and in one that had taken a standard verbal comprehension test under ordinary conditions.

Several studies have been concerned with the effects of feedback regarding test scores on the individual's subsequent test performance. In a particularly well-designed investigation with seventh-grade students, Bridgeman (1974) found that "success" feedback was followed by significantly higher performance on a similar test than was "failure" feedback in subjects who had actually performed equally well to begin with. This type of motivational feedback may operate largely through the goals the subjects set for themselves in subsequent performance and may thus represent another example of the self-fulfilling prophecy. Such general motivational feedback, however, should not be confused with corrective feedback, whereby the individual is informed about the specific items he missed and given remedial instruction; under these conditions, feedback is much more likely to improve the performance of initially low-scoring persons.

The examples cited in this section illustrate the wide diversity of test-related factors that may affect test scores. In the majority of well-administered testing programs, the influence of these factors is negligible for practical purposes. Nevertheless, the skilled examiner is constantly on guard to detect the possible operation of such factors and to minimize their influence. When circumstances do not permit the control of these conditions, the conclusions drawn from test performance should be qualified.

COACHING, PRACTICE, AND TEST SOPHISTICATION

In evaluating the effect of coaching or practice on test scores, a fundamental question is whether the improvement is limited to the specific items included in the test or whether it extends to the broader area of behavior that the test is designed to predict. The answer to this question represents the difference between coaching and education. Obviously any educational experience the individual undergoes, either formal or informal, in or out of school, should be reflected in his performance on tests sampling the relevant aspects of behavior. Such broad influences will in no way invalidate the test, since the test score presents an accurate picture of the individual's standing in the abilities under consideration. The difference is, of course, one of degree. Influences cannot be classified as either narrow or broad, but obviously vary widely in scope, from those affecting only a single administration of a single test, through those affecting performance on all items of a certain type, to those influencing the individual's performance in the large majority of his activities. From the standpoint of effective testing, however, a workable distinction can be made. Thus, it can be stated that a test score is invalidated only when a particular experience raises it without appreciably affecting the criterion behavior that the test is designed to predict.

COACHING. The effects of coaching on test scores have been widely investigated. Many of these studies were conducted by British psychologists, with special reference to the effects of practice and coaching on the tests formerly used in assigning 11-year-old children to different types of secondary schools (Yates et al., 1953–1954). As might be expected, the extent of improvement depends on the ability and earlier educational experiences of the examinees, the nature of the tests, and the amount and type of coaching provided. Individuals with deficient educational backgrounds are more likely to benefit from special coaching than are those who have had superior educational opportunities and are already prepared to do well on the tests. It is obvious, too, that the closer the resemblance between test content and coaching material, the greater will be the improvement in test scores. On the other hand, the more closely instruction is restricted to specific test content, the less likely is improvement to extend to criterion performance.

In America, the College Entrance Examination Board has been concerned about the spread of ill-advised commercial coaching courses for college applicants. To clarify the issues, the College Board conducted several well-controlled experiments to determine the effects of coaching on its Scholastic Aptitude Test and surveyed the results of similar studies by other, independent investigators (Angoff, 1971b; College Entrance Examination Board, 1968). These studies covered a variety of coaching methods and included students in both public and private high schools; one investigation was conducted with black students in 15 urban and rural high schools in Tennessee. The conclusion from all these studies is that intensive drill on items similar to those on the SAT is unlikely to produce appreciably greater gains than occur when students are retested with the SAT after a year of regular high school instruction.

On the basis of such research, the Trustees of the College Board issued a formal statement about coaching, in which the following points were made, among others (College Entrance Examination Board, 1968, pp. 8–9):

> The results of the coaching studies which have thus far been completed indicate that average increases of less than 10 points on a 600 point scale can be expected. It is not reasonable to believe that admissions decisions can be affected by such small changes in scores. This is especially true since the tests are merely supplementary to the school record and other evidence taken into account by admissions officers. . . . As the College Board uses the term, aptitude is not something fixed and impervious to influence by the way the child

lives and is taught. Rather, this particular Scholastic Aptitude Test is a measure of abilities that seem to grow slowly and stubbornly, profoundly influenced by conditions at home and at school over the years, but not responding to hasty attempts to relive a young lifetime.

It should also be noted that in its test construction procedures, the College Board investigates the susceptibility of new item types to coaching (Angoff, 1971b; Pike & Evans, 1972). Item types on which performance can be appreciably raised by short-term drill or instruction of a narrowly limited nature are not included in the operational forms of the tests.

PRACTICE. The effects of sheer repetition, or practice, on test performance are similar to the effects of coaching, but usually less pronounced. It should be noted that practice, as well as coaching, may alter the nature of the test, since the subjects may employ different work methods in solving the same problems. Moreover, certain types of items may be much easier when encountered a second time. An example is provided by problems requiring insightful solutions which, once attained, can be applied directly in solving the same or similar problems in a retest. Scores on such tests, whether derived from a repetition of the identical test or from a parallel form, should therefore be carefully scrutinized.

A number of studies have been concerned with the effects of the identical repetition of intelligence tests over periods ranging from a few days to several years (see Quereshi, 1968). Both adults and children, and both normal and mentally retarded persons have been employed. The studies have covered individual as well as group tests. All agree in showing significant mean gains on retests. Nor is improvement necessarily limited to the initial repetitions. Whether gains persist or level off in successive administrations seems to depend on the difficulty of the test and the ability level of the subjects. The implications of such findings are illustrated by the results obtained in annual retests of 3,500 schoolchildren with a variety of intelligence tests (Dearborn & Rothney, 1941). When the same test was readministered in successive years, the median IQ of the group rose from 102 to 113, but it dropped to 104 when another test was substituted. Because of the retest gains, the meaning of an IQ obtained on an initial and later trial proved to be quite different. For example, an IQ of 100 fell approximately at the average of the distribution on the initial trial, but in the lowest quarter on a retest. Such IQ's, though numerically identical and derived from the same test, might thus signify normal ability in the one instance and inferior ability in the other.

Gains in score are also found on retesting with parallel forms of the same test, although such gains tend in general to be smaller. Significant mean gains have been reported when alternate forms of a test were administered in immediate succession or after intervals ranging from one day to three years (Angoff, 1971b; Droege, 1966; Peel, 1951, 1952). Similar results have been obtained with normal and intellectually gifted schoolchildren, high school and college students, and employee samples. Data on the distribution of gains to be expected on a retest with a parallel form should be provided in test manuals and allowance for such gains should be made when interpreting test scores.

TEST SOPHISTICATION. The general problem of test sophistication should also be considered in this connection. The individual who has had extensive prior experience in taking psychological tests enjoys a certain advantage in test performance over one who is taking his first test (Heim & Wallace, 1949–1950; Millman, Bishop, & Ebel, 1965; Rodger, 1936). Part of this advantage stems from having overcome an initial feeling of strangeness, as well as from having developed more self-confidence and better test-taking attitudes. Part is the result of a certain amount of overlap in the type of content and functions covered by many tests. Specific familiarity with common item types and practice in the use of objective answer sheets may also improve performance slightly. It is particularly important to take test sophistication into account when comparing the scores obtained by children from different types of schools, where the extent of test-taking experience may have varied widely. Short orientation and practice sessions, as described earlier in this chapter, can be quite effective in equalizing test sophistication (Wahlstrom & Boersman, 1968).

CHAPTER 3

Social and Ethical Implications of Testing

IN ORDER to prevent the misuse of psychological tests, it has become necessary to erect a number of safeguards around both the tests themselves and the test scores. The distribution and use of psychological tests constitutes a major area in *Ethical Standards of Psychologists,* the code of professional ethics officially adopted by the American Psychological Association and reproduced in Appendix A. Principles 13, 14, and 15 are specifically directed to testing, being concerned with Test Security, Test Interpretation, and Test Publication. Other principles that, although broader in scope, are highly relevant to testing include 6 (Confidentiality), 7 (Client Welfare), and 9 (Impersonal Services). Some of the matters discussed in the *Ethical Standards* are closely related to points covered in the *Standards for Educational and Psychological Tests* (1974), cited in Chapter 1. For a fuller and richer understanding of the principles set forth in the *Ethical Standards,* the reader should consult two companion publications, the *Casebook on Ethical Standards of Psychologists* (1967) and *Ethical Principles in the Conduct of Research with Human Participants* (1973). Both report specific incidents to illustrate each principle. Special attention is given to marginal situations in which there may be a conflict of values, as between the advancement of science for human betterment and the protection of the rights and welfare of individuals.

USER QUALIFICATIONS

The requirement that tests be used only by appropriately qualified examiners is one step toward protecting the individual against the improper use of tests. Of course, the necessary qualifications vary with the type of test. Thus, a relatively long period of intensive training and supervised experience is required for the proper use of individual intelligence tests and most personality tests, whereas a minimum of specialized psychological training is needed in the case of educational achievement or vocational proficiency tests. It should also be noted that students who take tests in class for instructional purposes are not usually equipped to administer the tests to others or to interpret the scores properly.

The well-trained examiner chooses tests that are appropriate for both the particular purpose for which he is testing and the person to be examined. He is also cognizant of the available research literature on the chosen test and able to evaluate its technical merits with regard to such characteristics as norms, reliability, and validity. In administering the test, he is sensitive to the many conditions that may affect test performance, such as those illustrated in Chapter 2. He draws conclusions or makes recommendations only after considering the test score (or scores) in the light of other pertinent information about the individual. Above all, he should be sufficiently knowledgeable about the science of human behavior to guard against unwarranted inferences in his interpretations of test scores. When tests are administered by psychological technicians or assistants, or by persons in other professions, it is essential that an adequately qualified psychologist be available, at least as a consultant, to provide the needed perspective for a proper interpretation of test performance.

Misconceptions about the nature and purpose of tests and misinterpretations of test results underlie many of the popular criticisms of psychological tests. In part, these difficulties arise from inadequate communication between psychometricians and their various publics—educators, parents, legislators, job applicants, and so forth. Probably the most common examples center on unfounded inferences from IQs. Not all misconceptions about tests, however, can be attributed to inadequate communication between psychologists and laymen. Psychological testing itself has tended to become dissociated from the mainstream of behavioral science (Anastasi, 1967). The growing complexity of the science of psychology has inevitably been accompanied by increasing specialization among psychologists. In this process, psychometricians have concentrated more and more on the technical refinements of test construction and have tended to lose contact with developments in other relevant specialties, such as learning, child development, individual differences, and behavior genetics. Thus, the technical aspects of test construction have tended to outstrip the psychological sophistication with which test results are interpreted. Test scores can be properly interpreted only in the light of all available knowledge regarding the behavior that the tests are designed to measure.

Who is a qualified psychologist? Obviously, with the diversification of the field and the consequent specialization of training, no psychologist is equally qualified in all areas. In recognition of this fact, the *Ethical Standards* specify: "The psychologist recognizes the boundaries of his competence and the limitations of his techniques and does not offer services or use techniques that fail to meet professional standards established in particular fields" (Appendix A, Principle 2c). A useful distinction is that between a psychologist working in an institutional setting, such as a school system, university, clinic, or government agency, and one engaged in independent practice. Because the independent practitioner is less subject to judgment and evaluation by knowledgeable colleagues than is the institutional psychologist, he needs to meet higher standards of professional qualifications. The same would be true of a psychologist responsible for the supervision of other institutional psychologists or one who serves as an expert consultant to institutional personnel.

A significant step, both in upgrading professional standards and in helping the public to identify qualified psychologists, was the enactment of state licensing and certification laws for psychologists. Nearly all states now have such laws. Although the terms "licensing" and "certification" are often used interchangeably, in psychology certification typically refers to legal protection of the title "psychologist," whereas licensing controls the practice of psychology. Licensing laws thus need to include a definition of the practice of psychology. In either type of law, the requirements are generally a PhD in psychology, a specified amount of supervised experience, and satisfactory performance on a qualifying examination. Violations of the APA ethics code constitute grounds for revoking a certificate or license. Although most states began with the simpler certification laws, there has been continuing movement toward licensing.

At a more advanced level, specialty certification within psychology is provided by the American Board of Professional Psychology (ABPP). Requiring a high level of training and experience within designated specialties, ABPP grants diplomas in such areas as clinical, counseling, industrial and organizational, and school psychology. The Biographical Directory of the APA contains a list of current diplomates in each specialty, which can also be obtained directly from ABPP. The principal function of ABPP is to provide information regarding qualified psychologists. As a privately constituted board within the profession, ABPP does not have the enforcement authority available to the agencies administering the state licensing and certification laws.

TESTING INSTRUMENTS AND PROCEDURES

The purchase of tests is generally restricted to persons who meet certain minimal qualifications. The catalogues of major test publishers specify requirements that must be met by purchasers. Usually individuals with a master's degree in psychology or its equivalent qualify. Some publishers classify their tests into levels with reference to user qualifications, ranging from educational achievement and vocational proficiency tests, through group intelligence tests and interest inventories, to such clinical instruments as individual intelligence tests and most personality tests. Distinctions are also made between individual purchasers and authorized institutional purchasers of appropriate tests. Graduate students who may need a particular test for a class assignment or for research must have the purchase order countersigned by their psychology instructor, who assumes responsibility for the proper use of the test.

Efforts to restrict the distribution of tests have a dual objective: security of test materials and prevention of misuse. The *Ethical Standards* state: "Access to such devices is limited to persons with professional interests who will safeguard their use" (Principle 13); "Test scores, like test materials, are released only to persons who are qualified to interpret and use them properly" (Principle 14). It should be noted that although test distributors make sincere efforts to implement these objectives, the control they are able to exert is necessarily limited. The major responsibility for the proper use of tests resides in the individual user or institution concerned. It is evident, for example, that an MA degree in psychology —or even a PhD, state license, and ABPP diploma—do not necessarily signify that the individual is qualified to use a particular test or that his training is relevant to the proper interpretation of the results obtained with that test.

Another professional responsibility concerns the marketing of psychological tests by authors and publishers. Tests should not be released prematurely for general use. Nor should any claims be made regarding the merits of a test in the absence of sufficient objective evidence. When a test is distributed early for research purposes only, this condition should be clearly specified and the distribution of the test restricted accordingly. The test manual should provide adequate data to permit an evaluation of the test itself as well as full information regarding administration, scoring, and norms. The manual should be a factual exposition of what is known about the test rather than a selling device designed to put the test in a favorable light. It is the responsibility of the test author and publisher to revise tests and norms often enough to prevent obsolescence. The rapidity with which a test becomes outdated will, of course, vary widely with the nature of the test.

Tests or major parts of tests should not be published in a newspaper, magazine, or popular book, either for descriptive purposes or for self-evaluation. Under these conditions, self-evaluation would not only be subject to such drastic errors as to be well-nigh worthless, but it might also be psychologically injurious to the individual. Moreover, any publicity given to specific test items will tend to invalidate the future use of the test with other persons. It might also be added that presentation of test materials in this fashion tends to create an erroneous and distorted picture of psychological testing in general. Such publicity may foster either naïve credulity or indiscriminate resistance on the part of the public toward all psychological testing.

Another unprofessional practice is testing by mail. An individual's performance on either aptitude or personality tests cannot be properly assessed by mailing test forms to him and having him return them by mail for scoring and interpretation. Not only does this procedure provide no control of testing conditions but usually it also involves the interpretation of test scores in the absence of other pertinent information about the individual. Under these conditions, test results may be worse than useless.

PROTECTION OF PRIVACY

A question arising particularly in connection with personality tests is that of invasion of privacy. Insofar as some tests of emotional, motivational, or attitudinal traits are necessarily disguised, the subject may reveal characteristics in the course of such a test without realizing that he is so doing. Although there are few available tests whose approach is subtle enough to fall into this category, the possibility of developing such indirect testing procedures imposes a grave responsibility on the psychologist who uses them. For purposes of testing effectiveness it may be necessary to keep the examinee in ignorance of the specific ways in which his responses on any one test are to be interpreted. Nevertheless, a person should not be subjected to any testing program under false pretenses. Of primary importance in this connection is the obligation to have a clear understanding with the examinee regarding the use that will be made of his test results. The following statement contained in *Ethical Standards of Psychologists* (Principle 7d) is especially germane to this problem:

> The psychologist who asks that an individual reveal personal information in the course of interviewing, testing, or evaluation, or who allows such information to be divulged to him, does so only after making certain that the responsible person is fully aware of the purposes of the interview, testing, or evaluation and of the ways in which the information may be used.

Although concerns about the invasion of privacy have been expressed most commonly about personality tests, they logically apply to any type of test. Certainly any intelligence, aptitude, or achievement test may reveal limitations in skills and knowledge that an individual would rather not disclose. Moreover, any observation of an individual's behavior—as in an interview, casual conversation, or other personal encounter—may yield information about him that he would prefer to conceal and that he may reveal unwittingly. The fact that psychological tests have often been singled out in discussions of the invasion of privacy probably reflects prevalent misconceptions about tests. If all tests were recognized as measures of behavior samples, with no mysterious powers to penetrate beyond behavior, popular fears and suspicion would be lessened.

It should also be noted that all behavior research, whether employing tests or other observational procedures, presents the possibility of invasion of privacy. Yet, as scientists, psychologists are committed to the goal of advancing knowledge about human behavior. Principle 1a in *Ethical Standards of Psychologists* (Appendix A) clearly spells out the psychologist's conviction "that society will be best served when he investigates where his judgment indicates investigation is needed." Several other principles, on the other hand, are concerned with the protection of privacy and with the welfare of research subjects (see, e.g., 7d, 8a, 16). Conflicts of value may thus arise, which must be resolved in individual cases. Many examples of such conflict resolutions can be found in the previously cited *Ethical Principles in the Conduct of Research with Human Participants* (1973).

The problem is obviously not simple; and it has been the subject of extensive deliberation by psychologists and other professionals. In a report entitled *Privacy and Behavioral Research* (1967), prepared for the Office of Science and Technology, the right to privacy is defined as "the right of the individual to decide for himself how much he will share with others his thoughts, his feelings, and the facts of his personal life" (p. 2). It is further characterized as "a right that is essential to insure dignity and freedom of self-determination" (p. 2). To safeguard personal privacy, no universal rules can be formulated; only general guidelines can be provided. In the application of these guidelines to specific cases, there is no substitute for the ethical awareness and professional responsibility of the individual psychologist. Solutions must be worked out in terms of the particular circumstances.

One relevant factor is the purpose for which the testing is conducted—whether for individual counseling, institutional decisions regarding selection and classification, or research. In clinical or counseling situations, the client is usually willing to reveal himself in order to obtain help with his problems. The clinician or examiner does not invade privacy where he is freely admitted. Even under these conditions, however, the client should be warned that in the course of the testing or interviewing he may reveal information about himself without realizing that he is so doing; or he may disclose feelings of which he himself is unaware.

When testing is conducted for institutional purposes, the examinee should be fully informed as to the use that will be made of his test scores. It is also desirable, however, to explain to the examinee that correct assessment will benefit him, since it is not to his advantage to be placed in a position where he will fail or which he will find uncongenial. The results of tests administered in a clinical or counseling situation, of course, should not be made available for institutional purposes, unless the examinee gives his consent.

When tests are given for research purposes, anonymity should be preserved as fully as possible and the procedures for ensuring such anonymity should be explained in advance to the subjects. Anonymity does not, however, solve the problem of protecting privacy in all research contexts. Some subjects may resent the disclosure of facts they consider personal, even when complete confidentiality of responses is assured. In most cases, however, cooperation of subjects may be elicited if they are convinced that the information is needed for the research in question and if they have sufficient confidence in the integrity and competence of the investigator. All research on human behavior, whether or not it utilizes tests, may present conflicts of values. Freedom of inquiry, which is essential to the progress of science, must be balanced against the protection of the individual. The investigator must be alert to the values involved and must carefully weigh alternative solutions (see *Ethical Principles,* 1973; *Privacy and Behavioral Research,* 1967; Ruebhausen & Brim, 1966).

Whatever the purposes of testing, the protection of privacy involves two key concepts: relevance and informed consent. The information that the individual is asked to reveal must be *relevant* to the stated purposes of the testing. An important implication of this principle is that all practicable efforts should be made to ascertain the validity of tests for the particular diagnostic or predictive purpose for which they are used. An instrument that is demonstrably valid for a given purpose is one that provides relevant information. It also behooves the examiner to make sure that test scores are correctly interpreted. An individual is less likely to feel that his privacy is being invaded by a test assessing his readiness for a particular educational program than by a test allegedly measuring his "innate intelligence."

The concept of *informed consent* also requires clarification; and its application in individual cases may call for the exercise of considerable judgment (*Ethical Principles,* 1973; Ruebhausen & Brim, 1966). The examinee should certainly be informed about the purpose of testing, the kinds of data sought, and the use that will be made of his scores. It is not implied, however, that he be shown the test items in advance or told how specific responses will be scored. Nor should the test items be shown to a parent, in the case of a minor. Such information would usually invalidate the test. Not only would the giving of this information seriously impair the usefulness of an ability test, but it would also tend to distort responses on many personality tests. For example, if an individual is told in advance that a self-report inventory will be scored with a dominance scale, his responses are likely to be influenced by stereotyped (and often erroneous) ideas he may have about this trait, or by a false or distorted self-concept.

In the testing of children, special questions arise with regard to parental consent. Following an interdisciplinary conference, the Russell Sage Foundation (1970) published a set of *Guidelines for the Collection, Maintenance, and Dissemination of Pupil Records.* In reference to informed consent, the *Guidelines* differentiate between individual consent, given by the child, his parents, or both, and representational consent, given by the parents' legally elected or appointed representatives, such as a school board. While avoiding rigid prescriptions, the *Guidelines* suggest aptitude and achievement tests as examples of the type of instruments for which representational consent should be sufficient; at the other extreme, personality assessment is cited as requiring individual consent. A helpful feature of these *Guidelines* is the inclusion of sample letters and forms for obtaining written consent. There is also a selected bibliography on the ethical and legal aspects of school record keeping.

Testing procedures and experimental designs that protect the individual's right to decline to participate and that adequately safeguard his privacy, while yielding scientifically meaningful data, present a challenge to the psychologist's ingenuity. With proper rapport and the establishment of attitudes of mutual respect, however, the number of refusals to participate may be reduced to a negligible quantity. The technical difficulties of biased sampling and volunteer error may thus be avoided. Evidence from both national and statewide surveys suggests that this goal *can* be achieved, both in testing educational outcomes and in the more sensitive area of personality research (Holtzman, 1971; Womer, 1970). There is also some evidence that the number of respondents who feel that a personality inventory represents an invasion of privacy or who consider some of the items offensive is significantly reduced when the test is preceded by a simple and forthright explanation of how items were selected and how scores will be interpreted (Fink & Butcher, 1972). From the standpoint of test validity, it should be added that such an explanation did not affect the mean profile of scores on the personality inventories.

CONFIDENTIALITY

Like the protection of privacy, to which it is related, the problem of confidentiality of test data is multifaceted. The fundamental question is: Who shall have access to test results? Several considerations influence the answer in particular situations. Among them are the security of test content, the hazards of misunderstanding test scores, and the need of various persons to know the results.

There has been a growing awareness of the right of the individual himself to have access to the findings in his test report. He should also have the opportunity to comment on the contents of the report and if necessary to clarify or correct factual information. Counselors are now trying more and more to involve the client as an active participant in his own assessment. For these purposes, test results should be presented in a form that is readily understandable, free from technical jargon or labels, and oriented toward the immediate objective of the testing. Proper safeguards must be observed against misuse and misinterpretation of test findings (see *Ethical Standards*, Principle 14).

In the case of minors, one must also consider the parents' right of access to the child's test record. This presents a possible conflict with the child's own right to privacy, especially in the case of older children. In a searching analysis of the problem, Ruebhausen and Brim (1966, pp. 431–432) wrote: "Should not a child, even before the age of full legal responsibility, be accorded the dignity of a private personality? Considerations of healthy personal growth, buttressed with reasons of ethics, seem to command that this be done." The previously mentioned *Guidelines* (Russell Sage Foundation, 1970, p. 27) recommend that "when a student reaches the age of eighteen and no longer is attending high school, or is married (whether age eighteen or not)," he should have the right to deny parental access to his records. However, this recommendation is followed by the caution that school authorities check local state laws for possible legal difficulties in implementing such a policy.

Apart from these possible exceptions, the question is not *whether* to communicate test results to parents of a minor but *how* to do so. Parents normally have a legal right to information about their child; and it is usually desirable for them to have such information. In some cases, moreover, a child's academic or emotional difficulties may arise in part from parent-child relations. Under these conditions, the counselor's contact with the parents is of prime importance, both to fill in background data and to elicit parental cooperation.

Discussions of the confidentiality of test records have usually dealt with accessibility to a *third* person, other than the individual tested (or parent of a minor) and the examiner (*Ethical Standards*, Principle 6; Russell Sage Foundation, 1970). The underlying principle is that such records should *not* be released without the knowledge and consent of the individual.

When tests are administered in an institutional context, as in a school system, court, or employment setting, the individual should be informed at the time of testing regarding the purpose of the test, how the results will be used, and their availability to institutional personnel who have a legitimate need for them. Under these conditions, no further permission is required at the time results are made available within the institution. A different situation exists when test results are requested by outsiders, as when a prospective employer or a college requests test results from a school system. In these instances, individual consent for release of the data is required. The same requirement applies to tests administered in clinical and counseling contexts, or for research purposes. The previously cited *Guidelines* (Russell Sage Foundation, 1970, p. 42) contain a sample release form for the use of school systems in clearing the transmission of such data.

Another problem pertains to the retention of records in institutions. On the one hand, longitudinal records on individuals can be very valuable, not only for research purposes but also for understanding and counseling the person. As is so often the case, these advantages presuppose proper use and interpretation of test results. On the other hand, the availability of old records opens the way for such misuses as incorrect inferences from obsolete data and unauthorized access for other than the original testing purpose. It would be manifestly absurd, for example, to cite an IQ or a reading achievement score obtained by a child in the third grade when evaluating him for admission to college. Too much may have happened to him in the intervening years to make such early and isolated test scores meaningful. Similarly, when records are retained for many years, there is danger that they may be used for purposes that the individual (or his parents) never suspected and would not have approved.

To prevent such misuses, when records are retained either for legitimate longitudinal use in the interest of the individual or for acceptable research purposes, access to them should be subject to unusually stringent controls. In the *Guidelines* (Russell Sage Foundation, 1970, pp. 20–22), school records are classified into three categories with regard to their retention. A major determining factor in this classification is the degree of objectivity and verifiability of the data; another is relevance to the educational objectives of the school. It would be wise for any type of institution to formulate similar explicit policies regarding the destruction, retention, and accessibility of personal records.

The problems of maintenance, security, and accessibility of test results—and of all other personal data—have been magnified by the development of computerized data banks. In his preface to the *Guidelines* (Russell Sage Foundation, 1970, pp. 5–6), Ruebhausen wrote:

> Modern science has introduced a new dimension into the issues of privacy. There was a time when among the strongest allies of privacy were the inefficiency of man, the fallibility of his memory, and the healing compassion that accompanied both the passing of time and the warmth of human recollection. These allies are now being put to rout. Modern science has given us

the capacity to record faithfully, to maintain permanently, to retrieve promptly, and to communicate both widely and instantly.

The unprecedented advances in storing, processing, and retrieving data made possible by computers can be of inestimable service both in research and in the more immediate handling of social problems. The potential dangers of invasion of privacy and violation of confidentiality need to be faced squarely, constructively, and imaginatively. Rather than fearing the centralization and efficiency of complex computer systems, we should explore the possibility that these very characteristics may permit more effective procedures for protecting the security of individual records.

An example of what can be accomplished with adequate facilities is provided by the Link system developed by the American Council of Education (Astin & Boruch, 1970). In a longitudinal research program on the effects of different types of college environments, questionnaires were administered annually to several hundred thousand college freshmen. To permit the collection of follow-up data on the same persons while preventing the identification of individual responses by anyone at any future time, a three-file system of computer tapes was devised. The first tape, containing each student's responses marked with an arbitrary identification number, is readily accessible for research purposes. The second tape, containing only the students' names and addresses with the same identification numbers, was originally housed in a locked vault and used only to print labels for follow-up mailings. After the preparation of these tapes, the original questionnaires were destroyed.

This two-file system represents the traditional security system. It still did not provide complete protection, since some staff members would have access to both files. Moreover, such files are subject to judicial and legislative subpoena. For these reasons, a third file was prepared. Known as the Link file, it contained only the original identification numbers and a new set of random numbers which were substituted for the original identification numbers in the name and address file. The Link file was deposited at a computer facility in a foreign country, with the agreement that the file would never be released to anyone, including the American Council on Education. Follow-up data tapes are sent to the foreign facility, which substitutes one set of code numbers for the other. With the decoding files and the research data files under the control of different organizations, no one can identify the responses of individuals in the data files. Such elaborate precautions for the protection of confidentiality obviously would not be feasible *except* in a large-scale computerized data bank. The procedure could be simplified somewhat if the linking facility were located in a domestic agency given adequate protection against subpoena.

COMMUNICATING TEST RESULTS

Psychologists have given much thought to the communication of test results in a form that will be meaningful and useful. It is clear that the information should not be transmitted routinely, but should be accompanied by interpretive explanations by a professionally trained person. When communicating scores to parents, for example, a recommended procedure is to arrange a group meeting at which a counselor or school psychologist explains the purpose and nature of the tests, the sort of conclusions that may reasonably be drawn from the results, and the limitations of the data. Written reports about their own children may then be distributed to the parents, and arrangements made for personal interviews with any parents wishing to discuss the reports further. Regardless of how they are transmitted, however, an important condition is that test results should be presented in terms of descriptive performance levels rather than isolated numerical scores. This is especially true of intelligence tests, which are more likely to be misinterpreted than are achievement tests.

In communicating results to teachers, school administrators, employers, and other appropriate persons, similar safeguards should be provided. Broad levels of performance and qualitative descriptions in simple terms are to be preferred over specific numerical scores, except when communicating with adequately trained professionals. Even well-educated laymen have been known to confuse percentiles with percentage scores, percentiles with IQ's, norms with standards, and interest ratings with aptitude scores. But a more serious misinterpretation pertains to the conclusions drawn from test scores, even when their technical meaning is correctly understood. A familiar example is the popular assumption that an IQ indicates a fixed characteristic of the individual which predetermines his lifetime level of intellectual achievement.

In all test communication, it is desirable to take into account the characteristics of the person who is to receive the information. This applies not only to that person's general education and his knowledge about psychology and testing, but also to his anticipated emotional response to the information. In the case of a parent or teacher, for example, personal emotional involvement with the child may interfere with a calm and rational acceptance of factual information.

Last but by no means least is the problem of communicating test results to the individual himself, whether child or adult. The same general safeguards against misinterpretation apply here as in communicating with a third party. The person's emotional reaction to the information is especially important, of course, when he is learning about his own assets and shortcomings. When an individual is given his own test results, not only should the data be interpreted by a properly qualified person, but facilities should also be available for counseling anyone who may become emotionally disturbed by such information. For example, a college student might become seriously discouraged when he learns of his poor performance on a scholastic aptitude test. A gifted schoolchild might develop habits of laziness and shiftlessness, or he might become uncooperative and unmanageable, if he discovers that he is much brighter than any of his associates. A severe personality disorder may be precipitated when a maladjusted individual is given his score on a personality test. Such detrimental effects may, of course, occur regardless of the correctness or incorrectness of the score itself. Even when a test has been accurately administered and scored and properly interpreted, a knowledge of such a score without the opportunity to discuss it further may be harmful to the individual.

Counseling psychologists have been especially concerned with the development of effective ways of transmitting test information to their clients (see, e.g., Goldman, 1971, Ch. 14–16). Although the details of this process are beyond the scope of our present discussion, two major guidelines are of particular interest. First, test-reporting is to be viewed as an integral part of the counseling process and incorporated into the total counselor-client relationship. Second, insofar as possible, test results should be reported as answers to specific questions raised by the counselee. An important consideration in counseling relates to the counselee's acceptance of the information presented to him. The counseling situation is such that if the individual rejects any information, for whatever reasons, then that information is likely to be totally wasted.

TESTING AND THE CIVIL RIGHTS OF MINORITIES

THE SETTING. The decades since 1950 have witnessed an increasing public concern with the rights of minorities,[1] a concern that is reflected in the enactment of civil rights legislation at both federal and state levels. In connection with mechanisms for improving educational and vocational opportunities of such groups, psychological testing has been a major focus of attention. The psychological literature of the 1960s and early 1970s contains many discussions of the topic, whose impact ranges from clarification to obfuscation. Among the more clarifying contributions are several position papers by professional associations (see, e.g., American Psychological Association, 1969; Cleary, Humphreys, Kendrick, & Wesman, 1975; Deutsch, Fishman, Kogan, North, & Whiteman, 1964; *The responsible use of tests,* 1972). A brief but cogent paper by Flaugher (1974) also helps to clear away some prevalent sources of confusion.

---

[1] Although women represent a statistical majority in the national population, legally, occupationally, and in other ways they have shared many of the problems of minorities. Hence when the term "minority" is used in this section, it will be understood to include women.

Much of the concern centers on the lowering of test scores by cultural conditions that may have affected the development of aptitudes, interests, motivation, attitudes, and other psychological characteristics of minority group members. Some of the proposed solutions for the problem reflect misunderstandings about the nature and function of psychological tests. Differences in the experiential backgrounds of groups or individuals are inevitably manifested in test performance. Every psychological test measures a behavior sample. Insofar as culture affects behavior, its influence will and should be detected by tests. If we rule out all cultural differentials from a test, we may thereby lower its validity as a measure of the behavior domain it was designed to assess. In that case the test would fail to provide the kind of information needed to correct the very conditions that impaired performance.

Because the testing of minorities represents a special case within the broader problem of cross-cultural testing, the underlying theoretical rationale and testing procedures are discussed more fully in Chapter 12. A technical analysis of the concept of "test bias" is given in Chapter 7, in connection with test validity. In the present chapter, our interest is primarily in the basic issues and social implications of minority group testing.

TEST-RELATED FACTORS. In testing culturally diverse persons, it is important to differentiate between cultural factors that affect both test and criterion behavior and those whose influence is restricted to the test. It is the latter, test-related factors that reduce validity. Examples of such factors include previous experience in taking tests, motivation to perform well on tests, rapport with the examiner, and any other variables affecting performance on the particular test but irrelevant to the criterion under consideration. Special efforts should be made to reduce the operation of these test-related factors when testing persons with dissimilar cultural backgrounds. A desirable procedure is to provide adequate test-taking orientation and preliminary practice, as illustrated by the booklets and tape recordings cited in Chapter 2. Retesting with a parallel form is also recommended with low-scoring examinees who have had little or no prior test-taking experience.

Specific test content may also influence test scores in ways that are unrelated to criterion performance. In a test of arithmetic reasoning, for example, the use of names or pictures of objects unfamiliar in a particular cultural milieu would obviously represent a test-restricted handicap. Ability to carry out quantitative thinking does not depend upon familiarity with such objects. On the other hand, if the development of arithmetic ability itself is more strongly fostered in one culture than in another, scores on an arithmetic test should not eliminate or conceal such a difference.

Another, more subtle way in which specific test content may spuriously affect performance is through the examinee's emotional and attitudinal responses. Stories or pictures portraying typical suburban middle-class family scenes, for example, may alienate a child reared in a low-income inner-city home. Exclusive representation of the physical features of a single racial type in test illustrations may have a similar effect on members of an ethnic minority. In the same vein, women's organizations have objected to the perpetuation of sex stereotypes in test content, as in the portrayal of male doctors or executives and female nurses or secretaries. Certain words, too, may have acquired connotations that are offensive to minority groups. As one test publisher aptly expressed it, "Until fairly recently, most standardized tests were constructed by white middle-class people, who sometimes clumsily violate the feelings of the test-taker without even knowing it. In a way, one could say that we have been not so much culture biased as we have been 'culture blind'" (Fitzgibbon, 1972, pp. 2–3).

The major test publishers now make special efforts to weed out inappropriate test content. Their own test construction staffs have become sensitized to potentially offensive, culturally restricted, or stereotyped material. Members of different ethnic groups participate either as regular staff members or as consultants. And the reviewing of test content with reference to possible minority implications is a regular step in the process of test construction. An example of the application of these procedures in item construction and revision is provided by the 1970 edition of the Metropolitan Achievement Tests (Fitzgibbon, 1972; Harcourt Brace Jovanovich, 1972).

INTERPRETATION AND USE OF TEST SCORES. By far the most important considerations in the testing of culturally diverse groups—as in all testing—pertain to the interpretation of test scores. The most frequent misgivings regarding the use of tests with minority group members stem from misinterpretations of scores. If a minority examinee obtains a low score on an aptitude test or a deviant score on a personality test, it is essential to investigate why he did so. For example, an inferior score on an arithmetic test could result from low test-taking motivation, poor reading ability, or inadequate knowledge of arithmetic, among other reasons. Some thought should also be given to the type of norms to be employed in evaluating individual scores. Depending on the purpose of the testing, the appropriate norms may be general norms, subgroup norms based on persons with comparable experiential backgrounds, or the individual's own previous score.

In predicting outcome, such as performance on a job or in an educational program, we must also consider the effects of differential treatments. It is one of the contributions of decision theory to psychometrics that it provides ways of incorporating differential treatment into the prediction of outcomes from test scores (Ch. 7). For example, given certain test scores obtained by an individual with a specified cultural background, what will be his predicted college grades if we introduce remedial teaching programs, counseling designed to modify educational attitudes and motivation, or other appropriate treatments?

Tests are designed to show what an individual can do at a given point in time. They cannot tell us *why* he performs as he does. To answer that question, we need to investigate his background, motivations, and other pertinent circumstances. Nor can tests tell how able a culturally or educationally disadvantaged child might have been if he had been reared in a more favorable environment. Moreover, tests cannot compensate for cultural deprivation by eliminating its effects from their scores. On the contrary, tests should reveal such effects, so that appropriate remedial steps can be taken. To conceal the effects of cultural disadvantages by rejecting tests or by trying to devise tests that are insensitive to such effects can only retard progress toward a genuine solution of social problems. Such reactions toward tests are equivalent to breaking a thermometer because it registers a body temperature of 101°. Test scores should be used constructively: by the individual in enhancing self-understanding and personal development and in educational and vocational planning; by teachers in improving instruction and in adjusting instructional content to individual needs; and by employers in achieving a better match between persons and jobs, recognizing that persons can be trained and jobs can be redesigned.

The tendency to categorize and label, as a shortcut substitute for understanding, is all too prevalent. The diagnostic categories of classical psychiatry, whereby patients are assigned such labels as "paranoid schizophrenic" or "manic-depressive," are well known. Aware of the many deficiencies of this system of classification, clinical psychologists have turned increasingly to personality descriptions. Unlike the diagnostic labels, these descriptions focus on the origins and individual significance of deviant behavior and provide a more effective basis for therapy. But the traditional labels are not easily dislodged.

Another example of the categorizing tendency is provided by popular misinterpretations of the IQ. A common criticism of intelligence tests is that they encourage a rigid, inflexible, and permanent classification of individuals. A low IQ, it is argued, places an indelible stamp of inferiority on a child. In the case of culturally disadvantaged children, such an IQ would thus serve to perpetuate their handicap. It is largely because implications of permanent status have become attached to the IQ that in 1964 the use of group intelligence tests was discontinued in the New York City public schools (H. B. Gilbert, 1966; Loretan, 1966). That it proved necessary to discard the tests in order to eliminate the misconceptions about the fixity of the IQ is a revealing commentary on the tenacity of the misconceptions. It should also be noted that the use of individual intelligence tests like the Stanford-Binet, which are administered and interpreted by trained examiners and school psychologists, was not eliminated. It was the mass testing and routine use of IQs by relatively unsophisticated persons that was considered hazardous.

According to a popular misconception, the IQ is an index of innate intellectual potential and represents a fixed property of the organism. As will be seen in Chapter 12, this view is neither theoretically defensible nor supported by empirical data. When properly interpreted, intelligence test scores should not foster a rigid categorizing of persons. On the contrary, intelligence tests—and any other test—may be regarded as a map on which the individual's present position can be located. When combined with information about his experiential background, test scores should facilitate effective planning for the optimal development of the individual.

OBJECTIVITY OF TESTS. When social stereotypes and prejudice may distort interpersonal evaluations, tests provide a safeguard against favoritism and arbitrary or capricious decisions. Commenting on the use of tests in schools, Gardner (1961, pp. 48–49) wrote: "The tests couldn't see whether the youngster was in rags or in tweeds, and they couldn't hear the accents of the slum. The tests revealed intellectual gifts at every level of the population."

In the same vein, the *Guidelines for Testing Minority Group Children* (Deutsch et al., 1964, p. 139) contain the following observation:

> Many bright, non-conforming pupils, with backgrounds different from those of their teachers, make favorable showings on achievement tests, in contrast to their low classroom marks. These are very often children whose cultural handicaps are most evident in their overt social and interpersonal behavior. Without the intervention of standardized tests, many such children would be stigmatized by the adverse subjective ratings of teachers who tend to reward conformist behavior of middle-class character.

With regard to personnel selection, the contribution of tests was aptly characterized in the following words by John W. Macy, Jr., Chairman of the United States Civil Service Commission (*Testing and Public Policy*, 1965, p. 883):

The necessity to measure characteristics of people that are related to job performance is at the very root of the merit system, which is the basis for entry into the career services of the Federal Government. Thus, over the years, the public service has had a vital interest in the development and application of psychological testing methods. I have no doubt that the widespread public confidence in the objectivity of our hiring procedures has in large part been nurtured by the public's perception of the fairness, the practicality, and the integrity of the appraisal methods they must submit to.

The *Guidelines on Employee Selection Procedures,* prepared by the Equal Employment Opportunity Commission (1970) as an aid in the implementation of the Civil Rights Act, begin with the following statement of purpose:

> The guidelines in this part are based on the belief that properly validated and standardized employee selection procedures can significantly contribute to the implementation of nondiscriminatory personnel policies, as required by Title VII. It is also recognized that professionally developed tests, when used in conjunction with other tools of personnel assessment and complemented by sound programs of job design, may significantly aid in the development and maintenance of an efficient work force and, indeed, aid in the utilization and conservation of human resources generally.

In summary, tests *can* be misused in testing culturally disadvantaged persons—as in testing anyone else. When properly used, however, they serve an important function in preventing irrelevant and unfair discrimination. They also provide a quantitative index of the extent of cultural handicap as a necessary first step in remedial programs.

LEGAL REGULATIONS. A number of states enacted legislation and established Fair Employment Practices Commissions (FEPC) to implement it, prior to the development of such legal mechanisms at the federal level. Among the states that did so later, some efforts have been made to pattern the regulations after the federal model. The most pertinent federal legislation is provided by the Equal Employment Opportunity Act (Title VII of the Civil Rights Act of 1964 and its subsequent amendments).[2] Responsibility for implementation and enforcement is vested in the Equal Employment Opportunity Commission (EEOC). When charges are filed, the EEOC investigates the complaint and, if it finds the charges to be justified, tries first to correct the situation through conferences and voluntary compliance. If these procedures fail, EEOC may proceed to hold hearings, issue cease and desist orders, and finally bring action in the federal courts. In states having an approved FEPC, the Commission will defer to the local agency and will give its findings and conclusions "substantial weight."

---

[2] A brief summary of the major legal developments since midcentury, including legislative actions, executive orders, and court decisions, can be found in Fincher (1973).

The Office of Federal Contract Compliance (OFCC) has the authority to monitor the use of tests for employment purposes by government contractors. Colleges and universities are among the institutions concerned with OFCC regulations, because of their many research and training grants from such federal sources as the Department of Health, Education, and Welfare. Both EEOC and OFCC have drawn up guidelines regarding employee testing and other selection procedures, which are virtually identical in substance. A copy of the EEOC *Guidelines on Employee Selection Procedures* is reproduced in Appendix B, together with a 1974 amendment of the OFCC guidelines clarifying acceptable procedures for reporting test validity.[3]

Some major provisions in the EEOC *Guidelines* should be noted. The Equal Employment Opportunity Act prohibits discrimination by employers, trade unions, or employment agencies on the basis of race, color, religion, sex, or national origin. It is recognized that properly conducted testing programs not only are acceptable under this Act but can also contribute to the "implementation of nondiscriminatory personnel policies." Moreover, the same regulations specified for tests are also applied to all other formal and informal selection procedures, such as educational or work-history requirements, interviews, and application forms (Sections 2 and 13).

When the use of a test (or other selection procedure) results in a significantly higher rejection rate for minority candidates than for nonminority candidates, its utility must be justified by evidence of validity for the job in question. In defining acceptable procedures for establishing validity, the *Guidelines* make explicit reference to the *Standards for Educational and Psychological Tests* (1974) prepared by the American Psychological Association. A major portion of the *Guidelines* covers minimum requirements for acceptable validation (Sections 5 to 9). The reader may find it profitable to review these requirements after reading the more detailed technical discussion of validity in Chapters 6 and 7 of this book. It will be seen that the requirements are generally in line with good psychometric practice.

In the final section, dealing with affirmative action, the *Guidelines* point out that even when selection procedures have been satisfactorily validated, if disproportionate rejection rates result for minorities, steps must be taken to reduce this discrepancy as much as possible. Affirmative action implies that an organization does more than merely avoiding discriminatory practices. Psychologically, affirmative action programs may be regarded as efforts to compensate for the residual effects of past social inequities. Such effects may include deficiencies in aptitudes, job skills, attitudes, motivation, and other job-related behavior. They may also be reflected in a person's reluctance to apply for a job not traditionally open to minority candidates, or in his inexperience in job-seeking procedures. Examples of affirmative actions in meeting these problems include recruiting through media most likely to reach minorities; explicitly encouraging minority candidates to apply and following other recruiting practices designed to counteract past stereotypes; and, when practicable, providing special training programs for the acquisition of prerequisite skills and knowledge.

---

[3] In 1973, in the interest of simplification and improved coordination, the preparation of a set of uniform guidelines was undertaken by the Equal Employment Opportunity Coordinating Council, comprising representatives of EEOC, the Commission on Civil Rights, the Departments of Justice and Labor, and the Civil Service Commission. In 1976, a revision was adopted jointly by the last three agencies (*Federal Register*, Nov. 23, 1976, pp. 51734–51759).

PART 2

Principles of Psychological Testing

CHAPTER 4

Norms and the Interpretation of Test Scores

IN THE absence of additional interpretive data, a raw score on any psychological test is meaningless. To say that an individual has correctly solved 15 problems on an arithmetic reasoning test, or identified 34 words in a vocabulary test, or successfully assembled a mechanical object in 57 seconds conveys little or no information about his standing in any of these functions. Nor do the familiar percentage scores provide a satisfactory solution to the problem of interpreting test scores. A score of 65 percent correct on one vocabulary test, for example, might be equivalent to 30 percent correct on another, and to 80 percent correct on a third. The difficulty level of the items making up each test will, of course, determine the meaning of the score. Like all raw scores, percentage scores can be interpreted only in terms of a clearly defined and uniform frame of reference.

Scores on psychological tests are most commonly interpreted by reference to *norms* which represent the test performance of the standardization sample. The norms are thus empirically established by determining what a representative group of persons actually do on the test. Any individual's raw score is then referred to the distribution of scores obtained by the standardization sample, to discover where he falls in that distribution. Does his score coincide with the average performance of the standardization group? Is he slightly below average? Or does he fall near the upper end of the distribution?

In order to determine more precisely the individual's exact position with reference to the standardization sample, the raw score is converted into some relative measure. These derived scores are designed to serve a dual purpose. First, they indicate the individual's relative standing in the normative sample and thus permit an evaluation of his performance in reference to other persons. Second, they provide comparable measures that permit a direct comparison of the individual's performance on different tests. For example, if an individual has a raw score of 40 on a vocabulary test and a raw score of 22 on an arithmetic reasoning test, we obviously know nothing about his relative performance on the two tests. Is he better in vocabulary or in arithmetic, or equally good in both? Since raw scores on different tests are usually expressed in different units, a direct comparison of such scores is impossible. The difficulty level of the particular test would also affect such a comparison between raw scores. Derived scores, on the other hand, can be expressed in the same units and referred to the same or to closely similar normative samples for different tests. The individual's relative performance in many different functions can thus be compared.

There are various ways in which raw scores may be converted to fulfill the two objectives stated above. Fundamentally, however, derived scores are expressed in one of two major ways: (1) developmental level attained; (2) relative position within a specified group. These types of scores, together with some of their common variants, will be considered in separate sections of this chapter. But first it will be necessary to examine a few elementary statistical concepts that underlie the development and utilization of norms. The following section is included simply to clarify the meaning of certain common statistical measures. Simplified computational examples are given only for this purpose and not to provide training in statistical methods. For computational details and specific procedures to be followed in the practical application of these techniques, the reader is referred to any recent textbook on psychological or educational statistics.

STATISTICAL CONCEPTS

A major object of statistical method is to organize and summarize quantitative data in order to facilitate their understanding. A list of 1,000 test scores can be an overwhelming sight. In that form, it conveys little meaning. A first step in bringing order into such a chaos of raw data is to tabulate the scores into a *frequency distribution*, as illustrated in Table 1. Such a distribution is prepared by grouping the scores into convenient class intervals and tallying each score in the appropriate interval. When all scores have been entered, the tallies are counted to find the frequency, or number of cases, in each class interval. The sums of these frequencies will equal *N*, the total number of cases in the group. Table 1 shows the scores of 1,000 college students in a code-learning test in which one set of artificial words, or nonsense syllables, was to be substituted for another. The raw scores, giving number of correct syllables substituted during a two-minute trial, ranged from 8 to 52. They have been grouped into class intervals of 4 points, from 52–55 at the top of the distribution down to 8–11. The frequency column reveals that two persons scored

TABLE 1
Frequency Distribution of Scores of 1,000 College Students on a Code-Learning Test

(Data from Anastasi, 1934, p. 34)

Class Interval	Frequency
52–55	1
48–51	1
44–47	20
40–43	73
36–39	156
32–35	328
28–31	244
24–27	136
20–23	28
16–19	8
12–15	3
8–11	2
	1,000

between 8 and 11, three between 12 and 15, eight between 16 and 19, and so on.

The information provided by a frequency distribution can also be presented graphically in the form of a distribution curve. Figure 1 shows the data of Table 1 in graphic form. On the baseline, or horizontal axis, are the scores grouped into class intervals; on the vertical axis are the frequencies, or number of cases falling within each class interval. The graph has been plotted in two ways, both forms being in common use. In the *histogram,* the height of the column erected over each class interval corresponds to the number of persons scoring in that interval. We can think of each individual standing on another's shoulders to form the column. In the *frequency polygon,* the number of persons in each interval is indicated by a point placed in the center of the class interval and across from the appropriate frequency. The successive points are then joined by straight lines.

Except for minor irregularities, the distribution portrayed in Figure 1 resembles the bell-shaped *normal curve.* A mathematically determined, perfect normal curve is reproduced in Figure 3. This type of curve has important mathematical properties and provides the basis for many kinds of statistical analyses. For the present purpose, however, only a few features will be noted. Essentially, the curve indicates that the largest number of cases cluster in the center of the range and that the number drops off gradually in both directions as the extremes are approached. The curve is bilaterally symmetrical, with a single peak in the center. Most distributions of human traits, from height and weight to aptitudes and personality characteristics, approximate the normal curve. In general, the larger the group, the more closely will the distribution resemble the theoretical normal curve.

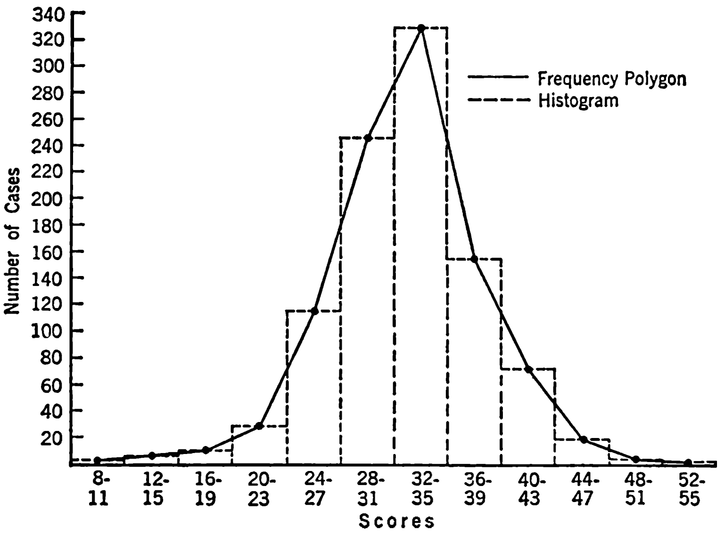


FIG. 1. Distribution Curves: Frequency Polygon and Histogram. (Data from Table 1.)

A group of scores can also be described in terms of some measure of *central tendency*. Such a measure provides a single, most typical or representative score to characterize the performance of the entire group. The most familiar of these measures is the average, more technically known as the *mean* (M). As is well known, this is found by adding all scores and dividing the sum by the number of cases (N). Another measure of central tendency is the *mode*, or most frequent score. In a frequency distribution, the mode is the midpoint of the class interval with the highest frequency. Thus, in Table 1, the mode falls midway between 32 and 35, being 33.5. It will be noted that this score corresponds to the highest point on the distribution curve in Figure 1. A third measure of central tendency is the *median*, or middlemost score when all scores have been arranged in order of size. The median is the point that bisects the distribution, half the cases falling above it and half below.

Further description of a set of test scores is given by measures of *variability*, or the extent of individual differences around the central tendency.

The most obvious and familiar way of reporting variability is in terms of the *range* between the highest and lowest score. The range, however, is extremely crude and unstable, for it is determined by only two scores. A single unusually high or low score would thus markedly affect its size. A more precise method of measuring variability is based on the difference between each individual's score and the mean of the group.

At this point it will be helpful to look at the example in Table 2, in which the various measures under consideration have been computed on 10 cases. Such a small group was chosen in order to simplify the demonstration, although in actual practice we would rarely perform these computations on so few cases. Table 2 serves also to introduce certain standard statistical symbols that should be noted for future reference. Original raw scores are conventionally designated by a capital X, and a small x is used to refer to deviations of each score from the group mean. The Greek letter Σ means "sum of." It will be seen that the first column in Table 2 gives the data for the computation of mean and median. The mean is 40; the median is 40.5, falling midway between 40 and 41—five cases

TABLE 2
Illustration of Central Tendency and Variability

	Score (X)	Diff. (x)		Diff. Squared (x^2)
50% of cases	48	+8	+20	64
	47	+7		49
	43	+3		9
	41	+1		1
	41	+1		1
Median = 40.5 →				
50% of cases	40	0	−20	0
	38	−2		4
	36	−4		16
	34	−6		36
	32	−8		64
	$\Sigma X = 400$	$\Sigma \lvert x \rvert = 40$		$\Sigma x^2 = 244$

$$M = \frac{\Sigma X}{N} = \frac{400}{10} = 40$$

$$AD = \frac{\Sigma \lvert x \rvert}{N} = \frac{40}{10} = 4$$

$$\text{Variance} = \sigma^2 = \frac{\Sigma x^2}{N} = \frac{244}{10} = 24.40$$

$$SD \text{ or } \sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{24.40} = 4.9$$

(50 percent) are above the median and five below. There is little point in finding a mode in such a small group, since the cases do not show clear-cut clustering on any one score. Technically, however, 41 would represent the mode, because two persons obtained this score, while all other scores occur only once.

The second column shows how far each score deviates above or below the mean of 40. The sum of these deviations will always equal zero, because the positive and negative deviations around the mean necessarily balance, or cancel each other out ($+20 - 20 = 0$). If we ignore signs, of course, we can average the absolute deviations, thus obtaining a measure known as the *average deviation* (*AD*). The symbol $|x|$ in the *AD* formula indicates that absolute values were summed, without regard to sign. Although of some descriptive value, the *AD* is not suitable for use in further mathematical analyses because of the arbitrary discarding of signs.

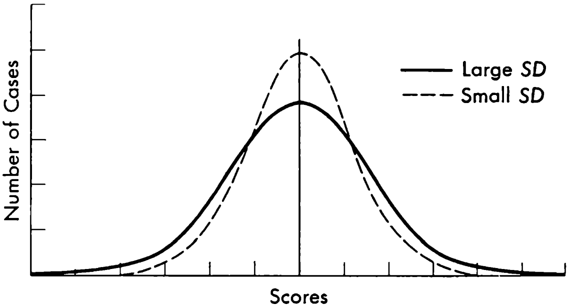


Fig. 2. Frequency Distributions with the Same Mean but Different Variability.

A much more serviceable measure of variability is the *standard deviation* (symbolized by either *SD* or σ), in which the negative signs are legitimately eliminated by squaring each deviation. This procedure has been followed in the last column of Table 2. The sum of this column divided by the number of cases $\left(\frac{\Sigma x^2}{N}\right)$ is known as the *variance*, or *mean square deviation*, and is symbolized by σ^2. The variance has proved extremely useful in sorting out the contributions of different factors to individual differences in test performance. For the present purposes, however, our chief concern is with the *SD*, which is the square root of the variance, as shown in Table 2. This measure is commonly employed in comparing the variability of different groups. In Figure 2, for example, are two distributions having the same mean but differing in variability. The distribution with wider individual differences yields a larger *SD* than the one with narrower individual differences.

The *SD* also provides the basis for expressing an individual's scores on

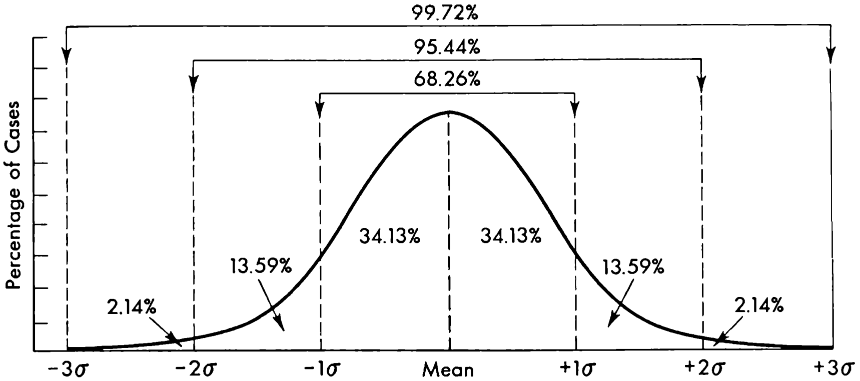


FIG. 3. Percentage Distribution of Cases in a Normal Curve.

different tests in terms of norms, as will be shown in the section on standard scores. The interpretation of the *SD* is especially clear-cut when applied to a normal or approximately normal distribution curve. In such a distribution, there is an exact relationship between the *SD* and the proportion of cases, as shown in Figure 3. On the baseline of this normal curve have been marked distances representing one, two, and three standard deviations above and below the mean. For instance, in the example given in Table 2, the mean would correspond to a score of 40, $+1\sigma$ to 44.9 $(40 + 4.9)$, $+2\sigma$ to 49.8 $(40 + 2 \times 4.9)$, and so on. The percentage of cases that fall between the mean and $+1\sigma$ in a normal curve is 34.13. Because the curve is symmetrical, 34.13 percent of the cases are likewise found between the mean and -1σ, so that between $+1\sigma$ and -1σ on both sides of the mean there are 68.26 percent of the cases. Nearly all the cases (99.72 percent) fall within $\pm 3\sigma$ from the mean. These relationships are particularly relevant in the interpretation of standard scores and percentiles, to be discussed in later sections.

DEVELOPMENTAL NORMS

One way in which meaning can be attached to test scores is to indicate how far along the normal developmental path the individual has progressed. Thus an 8-year-old who performs as well as the average 10-year-old on an intelligence test may be described as having a mental age of 10; a mentally retarded adult who performs at the same level would likewise be assigned an MA of 10. In a different context, a fourth-grade child may be characterized as reaching the sixth-grade norm in a reading test and the third-grade norm in an arithmetic test. Other developmental systems utilize more highly qualitative descriptions of behavior in specific functions, ranging from sensorimotor activities to concept formation. However expressed, scores based on developmental norms tend to be psychometrically crude and do not lend themselves well to precise statistical treatment. Nevertheless, they have considerable appeal for descriptive purposes, especially in the intensive clinical study of individuals and for certain research purposes.

MENTAL AGE. In Chapter 1 it was noted that the term "mental age" was widely popularized through the various translations and adaptations of the Binet-Simon scales, although Binet himself had employed the more neutral term "mental level." In age scales such as the Binet and its revisions, items are grouped into year levels. For example, those items passed by the majority of 7-year-olds in the standardization sample are placed in the 7-year level, those passed by the majority of 8-year-olds are assigned to the 8-year level, and so forth. A child's score on the test will then correspond to the highest year level that he can successfully complete. In actual practice, the individual's performance shows a certain amount of *scatter*. In other words, the subject fails some tests below his mental age level and passes some above it. For this reason, it is customary to compute the *basal age*, i.e., the highest age at and below which all tests are passed. Partial credits, in months, are then added to this basal age for all tests passed at higher year levels. The child's mental age on the test is the sum of the basal age and the additional months of credit earned at higher age levels.

Mental age norms have also been employed with tests that are not divided into year levels. In such a case, the subject's raw score is first determined. Such a score may be the total number of correct items on the whole test; or it may be based on time, on number of errors, or on some combination of such measures. The mean raw scores obtained by the children in each year group within the standardization sample constitute the age norms for such a test. The mean raw score of the 8-year-old children, for example, would represent the 8-year norm. If an individual's raw score is equal to the mean 8-year-old raw score, then his mental age on the test is 8 years. All raw scores on such a test can be transformed in a similar manner by reference to the age norms.

It should be noted that the mental age unit does not remain constant with age, but tends to shrink with advancing years. For example, a child who is one year retarded at age 4 will be approximately three years retarded at age 12. One year of mental growth from ages 3 to 4 is equivalent to three years of growth from ages 9 to 12. Since intellectual development progresses more rapidly at the earlier ages and gradually decreases as the individual approaches his mature limit, the mental age unit shrinks correspondingly with age. This relationship may be more readily visualized if we think of the individual's height as being expressed in terms of "height age." The difference, in inches, between a height age of 3 and 4 years would be greater than that between a height age of 10 and 11. Owing to the progressive shrinkage of the MA unit, one year of acceleration or retardation at, let us say, age 5 represents a larger deviation from the norm than does one year of acceleration or retardation at age 10.

GRADE EQUIVALENTS. Scores on educational achievement tests are often interpreted in terms of grade equivalents. This practice is understandable because the tests are employed within an academic setting. To describe a pupil's achievement as equivalent to seventh-grade performance in spelling, eighth-grade in reading, and fifth-grade in arithmetic has the same popular appeal as the use of mental age in the traditional intelligence tests.

Grade norms are found by computing the mean raw score obtained by children in each grade. Thus, if the average number of problems solved correctly on an arithmetic test by the fourth graders in the standardization sample is 23, then a raw score of 23 corresponds to a grade equivalent of 4. Intermediate grade equivalents, representing fractions of a grade, are usually found by interpolation, although they can also be obtained directly by testing children at different times within the school year. Because the school year covers ten months, successive months can be expressed as decimals. For example, 4.0 refers to average performance at the beginning of the fourth grade (September testing), 4.5 refers to average performance at the middle of the grade (February testing), and so forth.

Despite their popularity, grade norms have several shortcomings. First, the content of instruction varies somewhat from grade to grade. Hence, grade norms are appropriate only for common subjects taught throughout the grade levels covered by the test. They are not generally applicable at the high school level, where many subjects may be studied for only one or two years. Even with subjects taught in each grade, however, the emphasis placed on different subjects may vary from grade to grade, and progress may therefore be more rapid in one subject than in another during a particular grade. In other words, grade units are obviously unequal and these inequalities occur irregularly in different subjects.

Grade norms are also subject to misinterpretation unless the test user keeps firmly in mind the manner in which they were derived. For example, if a fourth-grade child obtains a grade equivalent of 6.9 in arithmetic, it does *not* mean that he has mastered the arithmetic processes taught in the sixth grade. He undoubtedly obtained his score largely by superior performance in fourth-grade arithmetic. It certainly could not be assumed that he has the prerequisites for seventh-grade arithmetic. Finally, grade norms tend to be incorrectly regarded as performance standards. A sixth-grade teacher, for example, may assume that all pupils in her class should fall at or close to the sixth-grade norm in achievement tests. This misconception is certainly not surprising when grade norms are used. Yet individual differences within any one grade are such that the range of achievement test scores will inevitably extend over several grades.

ORDINAL SCALES. Another approach to developmental norms derives from research in child psychology. Empirical observation of behavior development in infants and young children led to the description of behavior typical of successive ages in such functions as locomotion, sensory discrimination, linguistic communication, and concept formation. An early example is provided by the work of Gesell and his associates at Yale (Ames, 1937; Gesell et al., 1940; Gesell & Amatruda, 1947; Halverson, 1933). The Gesell Developmental Schedules show the approximate developmental level in months that the child has attained in each of four major areas of behavior, namely, motor, adaptive, language, and personal-social. These levels are found by comparing the child's behavior with that typical of eight key ages, ranging from 4 weeks to 36 months.

Gesell and his co-workers emphasized the sequential patterning of early behavior development. They cited extensive evidence of uniformities of developmental sequences and an orderly progression of behavior changes. For example, the child's reactions toward a small object placed in front of him exhibit a characteristic chronological sequence in visual fixation and in hand and finger movements. Use of the entire hand in crude attempts at palmar prehension occurs at an earlier age than use of the thumb in opposition to the palm; this type of prehension is in turn followed by use of the thumb and index finger in a more efficient pincer-like grasp of the object. Such sequential patterning was likewise observed in walking, stair climbing, and most of the sensorimotor development of the first few years. The scales developed within this framework are ordinal in the sense that developmental stages follow in a constant order, each stage presupposing mastery of prerequisite behavior characteristic of earlier stages.[1]

---

[1] This usage of the term "ordinal scale" differs from that in statistics, in which an ordinal scale is simply one that permits a rank-ordering of individuals without knowledge about amount of difference between them; in the statistical sense, ordinal scales are contrasted to equal-unit interval scales. Ordinal scales of child development are actually designed on the model of a Guttman scale or simplex, in which successful performance at one level implies success at all lower levels (Guttman, 1944). An

Since the 1960s, there has been a sharp upsurge of interest in the developmental theories of the Swiss child psychologist, Jean Piaget (see Flavell, 1963; Ginsburg & Opper, 1969; Green, Ford, & Flamer, 1971). Piaget's research has focused on the development of cognitive processes from infancy to the midteens. He is concerned with specific concepts rather than broad abilities. An example of such a concept, or schema, is object permanence, whereby the child is aware of the identity and continuing existence of objects when they are seen from different angles or are out of sight. Another widely studied concept is conservation, or the recognition that an attribute remains constant over changes in perceptual appearance, as when the same quantity of liquid is poured into differently shaped containers, or when rods of the same length are placed in different spatial arrangements.

Piagetian tasks have been used widely in research by developmental psychologists and some have been organized into standardized scales, to be discussed in Chapters 10 and 14 (Goldschmid & Bentler, 1968b; Loretan, 1966; Pinard & Laurendeau, 1964; Uzgiris & Hunt, 1975). In accordance with Piaget's approach, these instruments are ordinal scales, in which the attainment of one stage is contingent upon completion of the earlier stages in the development of the concept. The tasks are designed to reveal the dominant aspects of each developmental stage; only later are empirical data gathered regarding the ages at which each stage is typically reached. In this respect, the procedure differs from that followed in constructing age scales, in which items are selected in the first place on the basis of their differentiating between successive ages.

In summary, ordinal scales are designed to identify the stage reached by the child in the development of specific behavior functions. Although scores may be reported in terms of approximate age levels, such scores are secondary to a qualitative description of the child's characteristic behavior. The ordinality of such scales refers to the uniform progression of development through successive stages. Insofar as these scales typically provide information about what the child is actually able to do (e.g., climbs stairs without assistance; recognizes identity in quantity of liquid when poured into differently shaped containers), they share important features with the criterion-referenced tests to be discussed in a later section of this chapter.

WITHIN-GROUP NORMS

Nearly all standardized tests now provide some form of within-group norms. With such norms, the individual's performance is evaluated in

---

extension of Guttman's analysis to include nonlinear hierarchies is described by Bart and Airasian (1974), with special reference to Piagetian scales.

terms of the performance of the most nearly comparable standardization group, as when comparing a child's raw score with that of children of the same chronological age or in the same school grade. Within-group scores have a uniform and clearly defined quantitative meaning and can be appropriately employed in most types of statistical analysis.

PERCENTILES. Percentile scores are expressed in terms of the percentage of persons in the standardization sample who fall below a given raw score. For example, if 28 percent of the persons obtain fewer than 15 problems correct on an arithmetic reasoning test, then a raw score of 15 corresponds to the 28th percentile (P_{28}). A percentile indicates the individual's relative position in the standardization sample. Percentiles can also be regarded as ranks in a group of 100, except that in ranking it is customary to start counting at the top, the best person in the group receiving a rank of one. With percentiles, on the other hand, we begin counting at the bottom, so that the lower the percentile, the poorer the individual's standing.

The 50th percentile (P_{50}) corresponds to the median, already discussed as a measure of central tendency. Percentiles above 50 represent above-average performance; those below 50 signify inferior performance. The 25th and 75th percentile are known as the first and third quartile points (Q_1 and Q_3), because they cut off the lowest and highest quarters of the distribution. Like the median, they provide convenient landmarks for describing a distribution of scores and comparing it with other distributions.

Percentiles should not be confused with the familiar percentage scores. The latter are raw scores, expressed in terms of the percentage of correct items; percentiles are derived scores, expressed in terms of percentage of persons. A raw score lower than any obtained in the standardization sample would have a percentile rank of zero (P_0); one higher than any score in the standardization sample would have a percentile rank of 100 (P_{100}). These percentiles, however, do not imply a zero raw score and a perfect raw score.

Percentile scores have several advantages. They are easy to compute and can be readily understood, even by relatively untrained persons. Moreover, percentiles are universally applicable. They can be used equally well with adults and children and are suitable for any type of test, whether it measures aptitude or personality variables.

The chief drawback of percentile scores arises from the marked inequality of their units, especially at the extremes of the distribution. If the distribution of raw scores approximates the normal curve, as is true of most test scores, then raw score differences near the median or center of the distribution are exaggerated in the percentile transformation, whereas raw score differences near the ends of the distribution are greatly shrunk. This distortion of distances between scores can be seen in Figure 4. In a normal curve, it will be recalled, cases cluster closely at the center and scatter more widely as the extremes are approached. Consequently, any given percentage of cases near the center covers a shorter distance on the baseline than the same percentage near the ends of the distribution. In Figure 4, this discrepancy in the gaps between percentile ranks (*PR*) can readily be seen if we compare the distance between a *PR* of 40 and a *PR* of 50 with that between a *PR* of 10 and a *PR* of 20. Even more striking is the discrepancy between these distances and that between a *PR* of 10 and *PR* of 1. (In a mathematically derived normal curve, zero percentile is not reached until infinity and hence cannot be shown on the graph.)

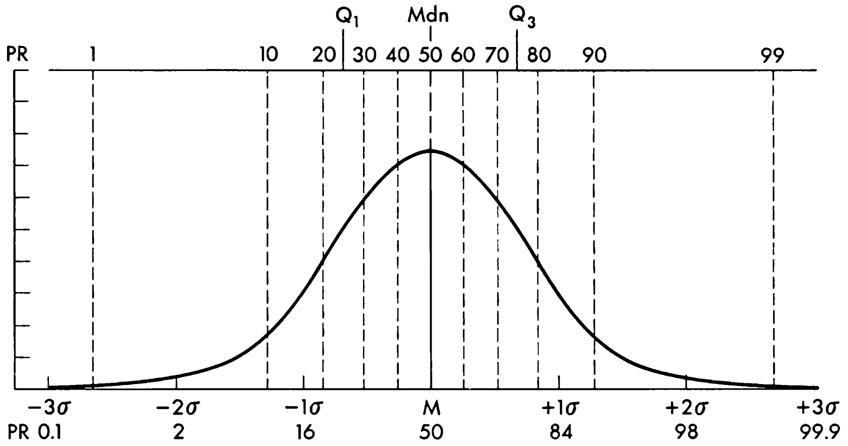


FIG. 4. Percentile Ranks in a Normal Distribution.

The same relationship can be seen from the opposite direction if we examine the percentile ranks corresponding to equal σ-distances from the mean of a normal curve. These percentile ranks are given under the graph in Figure 4. Thus, the percentile difference between the mean and +1σ is 34 (84 − 50). That between +1σ and +2σ is only 14 (98 − 84).

It is apparent that percentiles show each individual's relative position in the normative sample but not the amount of difference between scores. If plotted on arithmetic probability paper, however, percentile scores can also provide a correct visual picture of the differences between scores. Arithmetic probability paper is a cross-section paper in which the vertical lines are spaced in the same way as the percentile points in a normal distribution (as in Figure 4), whereas the horizontal lines are uniformly spaced, or vice versa (as in Figure 5). Such *normal percentile*

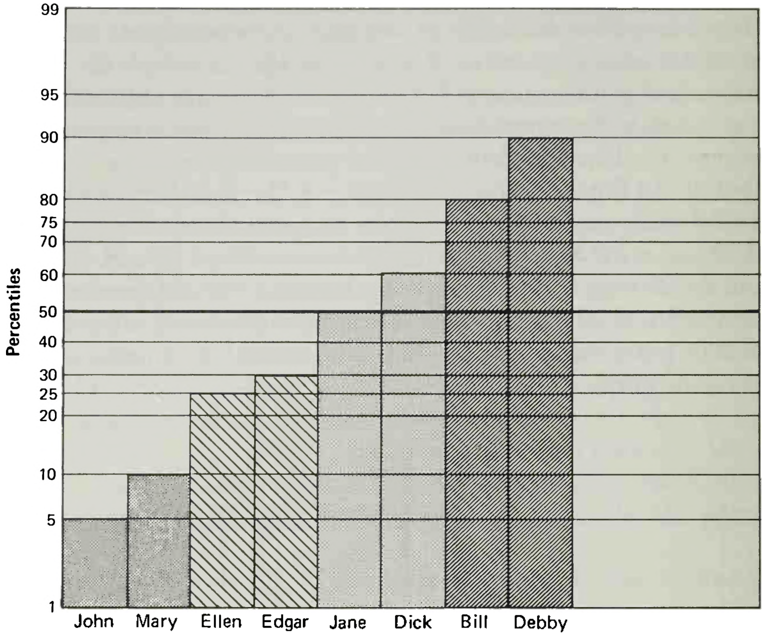


FIG. 5. A Normal Percentile Chart. Percentiles are spaced so as to correspond to equal distances in a normal distribution. Compare the score distance between John and Mary with that between Ellen and Edgar; within both pairs, the percentile difference is 5 points. Jane and Dick differ by 10 percentile points, as do Bill and Debby.

charts can be used to plot the scores of different persons on the same test or the scores of the same person on different tests. In either case, the actual interscore difference will be correctly represented. Many aptitude and achievement batteries now utilize this technique in their score profiles, which show the individual's performance in each test. An example is the Individual Report Form of the Differential Aptitude Tests, reproduced in Figure 13 (Ch. 5).

STANDARD SCORES. Current tests are making increasing use of standard scores, which are the most satisfactory type of derived score from most points of view. Standard scores express the individual's distance from the mean in terms of the standard deviation of the distribution.

Standard scores may be obtained by either linear or nonlinear transformations of the original raw scores. When found by a *linear transformation,* they retain the exact numerical relations of the original raw scores, because they are computed by subtracting a constant from each raw score and then dividing the result by another constant. The relative magnitude of differences between standard scores derived by such a linear transformation corresponds exactly to that between the raw scores. All properties of the original distribution of raw scores are duplicated in the distribution of these standard scores. For this reason, any computations that can be carried out with the original raw scores can also be carried out with linear standard scores, without any distortion of results.

Linearly derived standard scores are often designated simply as "standard scores" or "z scores." To compute a z score, we find the difference between the individual's raw score and the mean of the normative group and then divide this difference by the *SD* of the normative group. Table 3 shows the computation of z scores for two individuals, one of whom falls 1 *SD* above the group mean, the other .40 *SD* below the mean. Any raw score that is exactly equal to the mean is equivalent to a z score of zero. It is apparent that such a procedure will yield derived scores that have a negative sign for all subjects falling below the mean. Moreover, because the total range of most groups extends no farther than about 3 *SD*'s above and below the mean, such standard scores will have to be reported to at least one decimal place in order to provide sufficient differentiation among individuals.

Table 3
Computation of Standard Scores

$$z = \frac{X - M}{SD} \qquad M = 60 \qquad SD = 5$$

JOHN'S SCORE	*BILL'S SCORE*
$X_1 = 65$	$X_2 = 58$
$z_1 = \frac{65 - 60}{5}$	$z_2 = \frac{58 - 60}{5}$
$= +1.00$	$= -0.40$

Both the above conditions, viz., the occurrence of negative values and of decimals, tend to produce awkward numbers that are confusing and difficult to use for both computational and reporting purposes. For this reason, some further linear transformation is usually applied, simply to put the scores into a more convenient form. For example, the scores on the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board are standard scores adjusted to a mean of 500 and an *SD* of 100. Thus a standard score of -1 on this test would be expressed as 400 ($500 - 100 = 400$). Similarly, a standard score of $+1.5$ would correspond to 650 ($500 + 1.5 \times 100 = 650$). To convert an original standard score to the new scale, it is simply necessary to multiply the standard score by the desired *SD* (100) and add it to or subtract it from the desired mean (500). Any other convenient values can be arbitrarily chosen for the new mean and *SD*. Scores on the separate subtests of the Wechsler Intelligence Scales, for instance, are converted to a distribution with a mean of 10 and an *SD* of 3. All such measures are examples of linearly transformed standard scores.

It will be recalled that one of the reasons for transforming raw scores into any derived scale is to render scores on different tests comparable. The linearly derived standard scores discussed in the preceding section will be comparable only when found from distributions that have approximately the same form. Under such conditions, a score corresponding to 1 *SD* above the mean, for example, signifies that the individual occupies the same position in relation to both groups. His score exceeds approximately the same percentage of persons in both distributions, and this percentage can be determined if the form of the distribution is known. If, however, one distribution is markedly skewed and the other normal, a z score of +1.00 might exceed only 50 percent of the cases in one group but would exceed 84 percent in the other.

In order to achieve comparability of scores from dissimilarly shaped distributions, nonlinear transformations may be employed to fit the scores to any specified type of distribution curve. The mental age and percentile scores described in earlier sections represent nonlinear transformations, but they are subject to other limitations already discussed. Although under certain circumstances another type of distribution may be more appropriate, the normal curve is usually employed for this purpose. One of the chief reasons for this choice is that most raw score distributions approximate the normal curve more closely than they do any other type of curve. Moreover, physical measures such as height and weight, which use equal-unit scales derived through physical operations, generally yield normal distributions.[2] Another important advantage of the normal curve is that it has many useful mathematical properties, which facilitate further computations.

Normalized standard scores are standard scores expressed in terms of a distribution that has been transformed to fit a normal curve. Such scores can be computed by reference to tables giving the percentage of cases falling at different *SD* distances from the mean of a normal curve. First, the percentage of persons in the standardization sample falling at or above each raw score is found. This percentage is then located in the normal curve frequency table, and the corresponding normalized standard score is obtained. Normalized standard scores are expressed in the same form as linearly derived standard scores, viz., with a mean of zero and an *SD* of 1. Thus, a normalized score of zero indicates that the individual falls at the mean of a normal curve, excelling 50 percent of the group. A score of −1 means that he surpasses approximately 16 percent of the group; and a score of +1, that he surpasses 84 percent. These percentages correspond to a distance of 1 *SD* below and 1 *SD* above the mean of a normal curve, respectively, as can be seen by reference to the bottom line of Figure 4.

---

[2] Partly for this reason and partly as a result of other theoretical considerations, it has frequently been argued that, by normalizing raw scores, an equal-unit scale could be developed for psychological measurement similar to the equal-unit scales of physical measurement. This, however, is a debatable point that involves certain questionable assumptions.

Like linearly derived standard scores, normalized standard scores can be put into any convenient form. If the normalized standard score is multiplied by 10 and added to or subtracted from 50, it is converted into a *T score,* a type of score first proposed by McCall (1922). On this scale, a score of 50 corresponds to the mean, a score of 60 to 1 *SD* above the mean, and so forth. Another well-known transformation is represented by the *stanine* scale, developed by the United States Air Force during World War II. This scale provides a single-digit system of scores with a mean of 5 and an *SD* of approximately 2.[3] The name *stanine* (a contraction of "*sta*ndard *nine*") is based on the fact that the scores run from 1 to 9. The restriction of scores to single-digit numbers has certain computational advantages, for each score requires only a single column on computer punched cards.

TABLE 4
Normal Curve Percentages for Use in Stanine Conversion

Percentage	4	7	12	17	20	17	12	7	4
Stanine	1	2	3	4	5	6	7	8	9

Raw scores can readily be converted to stanines by arranging the original scores in order of size and then assigning stanines in accordance with the normal curve percentages reproduced in Table 4. For example, if the group consists of exactly 100 persons, the 4 lowest-scoring persons receive a stanine score of 1, the next 7 a score of 2, the next 12 a score of 3, and so on. When the group contains more or fewer than 100 cases, the number corresponding to each designated percentage is first computed, and these numbers of cases are then given the appropriate stanines.

---

[3] Kaiser (1958) proposed a modification of the stanine scale that involves slight changes in the percentages and yields an *SD* of exactly 2, thus being easier to handle quantitatively. Other variants are the *C* scale (Guilford & Fruchter, 1973, Ch. 19), consisting of 11 units and also yielding an *SD* of 2, and the 10-unit *sten* scale, with 5 units above and 5 below the mean (Canfield, 1951).

Thus, out of 200 cases, 8 would be assigned a stanine of 1 (4 percent of 200 = 8). With 150 cases, 6 would receive a stanine of 1 (4 percent of 150 = 6). For any group containing from 10 to 100 cases, Bartlett and Edgerton (1966) have prepared a table whereby ranks can be directly converted to stanines. Because of their practical as well as theoretical advantages, stanines are being used increasingly, especially with aptitude and achievement tests.

Although normalized standard scores are the most satisfactory type of score for the majority of purposes, there are nevertheless certain technical objections to normalizing all distributions routinely. Such a transformation should be carried out only when the sample is large and representative and when there is reason to believe that the deviation from normality results from defects in the test rather than from characteristics of the sample or from other factors affecting the behavior under consideration. It should also be noted that when the original distribution of raw scores approximates normality, the linearly derived standard scores and the normalized standard scores will be very similar. Although the methods of deriving these two types of scores are quite different, the resulting scores will be nearly identical under such conditions. Obviously, the process of normalizing a distribution that is already virtually normal will produce little or no change. Whenever feasible, it is generally more desirable to obtain a normal distribution of raw scores by proper adjustment of the difficulty level of test items rather than by subsequently normalizing a markedly nonnormal distribution. With an approximately normal distribution of raw scores, the linearly derived standard scores will serve the same purposes as normalized standard scores.

THE DEVIATION IQ. In an effort to convert MA scores into a uniform index of the individual's relative status, the ratio IQ (Intelligence Quotient) was introduced in early intelligence tests. Such an IQ was simply the ratio of mental age to chronological age, multiplied by 100 to eliminate decimals (IQ = 100 × MA/CA). Obviously, if a child's MA equals his CA, his IQ will be exactly 100. An IQ of 100 thus represents normal or average performance. IQ's below 100 indicate retardation, those above 100, acceleration.

The apparent logical simplicity of the traditional ratio IQ, however, soon proved deceptive. A major technical difficulty is that, unless the *SD* of the IQ distribution remains approximately constant with age, IQ's will not be comparable at different age levels. An IQ of 115 at age 10, for example, may indicate the same degree of superiority as an IQ of 125 at age 12, since both may fall at a distance of 1 *SD* from the means of their respective age distributions. In actual practice, it proved very difficult to construct tests that met the psychometric requirements for comparability of ratio IQ's throughout their age range. Chiefly for this reason, the ratio IQ has been largely replaced by the so-called deviation IQ, which is actually another variant of the familiar standard score. The deviation IQ is a standard score with a mean of 100 and an *SD* that approximates the *SD* of the Stanford-Binet IQ distribution. Although the *SD* of the Stanford-Binet ratio IQ (last used in the 1937 edition) was not exactly constant at all ages, it fluctuated around a median value slightly greater than 16. Hence, if an *SD* close to 16 is chosen in reporting standard scores on a newly developed test, the resulting scores can be interpreted in the same way as Stanford-Binet ratio IQ's. Since Stanford-Binet IQ's have been in use for many years, testers and clinicians have become accustomed to interpreting and classifying test performance in terms of such IQ levels. They have learned what to expect from individuals with IQ's of 40, 70, 90, 130, and so forth. There are therefore certain practical advantages in the use of a derived scale that corresponds to the familiar distribution of Stanford-Binet IQ's. Such a correspondence of score units can be achieved by the selection of numerical values for the mean and *SD* that agree closely with those in the Stanford-Binet distribution.

It should be added that the use of the term "IQ" to designate such standard scores may seem to be somewhat misleading. Such IQ's are not derived by the same methods employed in finding traditional ratio IQ's. They are not ratios of mental ages and chronological ages. The justification lies in the general familiarity of the term "IQ," and in the fact that such scores *can* be interpreted as IQ's provided that their *SD* is approximately equal to that of previously known IQ's. Among the first tests to express scores in terms of deviation IQ's were the Wechsler Intelligence Scales. In these tests, the mean is 100 and the *SD* 15. Deviation IQ's are also used in a number of current group tests of intelligence and in the latest revision of the Stanford-Binet itself.

With the increasing use of deviation IQ's, it is important to remember that deviation IQ's from different tests are comparable only when they employ the same or closely similar values for the *SD*. This value should always be reported in the manual and carefully noted by the test user. If a test maker chooses a different value for the *SD* in making up his deviation IQ scale, the meaning of any given IQ on his test will be quite different from its meaning on other tests. These discrepancies are illustrated in Table 5, which shows the percentage of cases in normal distributions with *SD*'s from 12 to 18 who would obtain IQ's at different levels. These *SD* values have actually been employed in the IQ scales of published tests. Table 5 shows, for example, that an IQ of 70 cuts off the lowest 3.1 percent when the *SD* is 16 (as in the Stanford-Binet); but it may cut off as few as 0.7 percent ($SD = 12$) or as many as 5.1 percent ($SD = 18$). An IQ of 70 has been used traditionally as a cutoff point for identifying

TABLE 5
Percentage of Cases at Each IQ Interval in Normal Distributions with Mean of 100 and Different Standard Deviations

(Courtesy Test Department, Harcourt Brace Jovanovich, Inc.)

IQ Interval	Percentage Frequency			
	SD = 12	SD = 14	SD = 16	SD = 18
130 and above	0.7	1.6	3.1	5.1
120–129	4.3	6.3	7.5	8.5
110–119	15.2	16.0	15.8	15.4
100–109	29.8 } 59.6	26.1 } 52.2	23.6 } 47.2	21.0 } 42.0
90– 99	29.8	26.1	23.6	21.0
80– 89	15.2	16.0	15.8	15.4
70– 79	4.3	6.3	7.5	8.5
Below 70	0.7	1.6	3.1	5.1
Total	100.0	100.0	100.0	100.0

mental retardation. The same discrepancies, of course, apply to IQ's of 130 and above, which might be used in selecting children for special programs for the intellectually gifted. The IQ range between 90 and 110, generally described as normal, may include as few as 42 percent or as many as 59.6 percent of the population, depending on the test chosen. To be sure, test publishers are making efforts to adopt the uniform *SD* of 16 in new tests and in new editions of earlier tests. There are still enough variations among currently available tests, however, to make the checking of the *SD* imperative.

INTERRELATIONSHIPS OF WITHIN-GROUP SCORES. At this stage in our discussion of derived scores, the reader may have become aware of a rapprochement among the various types of scores. Percentiles have gradually been taking on at least a graphic resemblance to normalized standard scores. Linear standard scores are indistinguishable from normalized standard scores if the original distribution of raw scores closely approximates the normal curve. Finally, standard scores have become IQ's and vice versa. In connection with the last point, a reexamination of the meaning of a ratio IQ on such a test as the Stanford-Binet will show that these IQ's can themselves be interpreted as standard scores. If we know that the distribution of Stanford-Binet ratio IQ's had a mean of 100 and an *SD* of approximately 16, we can conclude that an IQ of 116

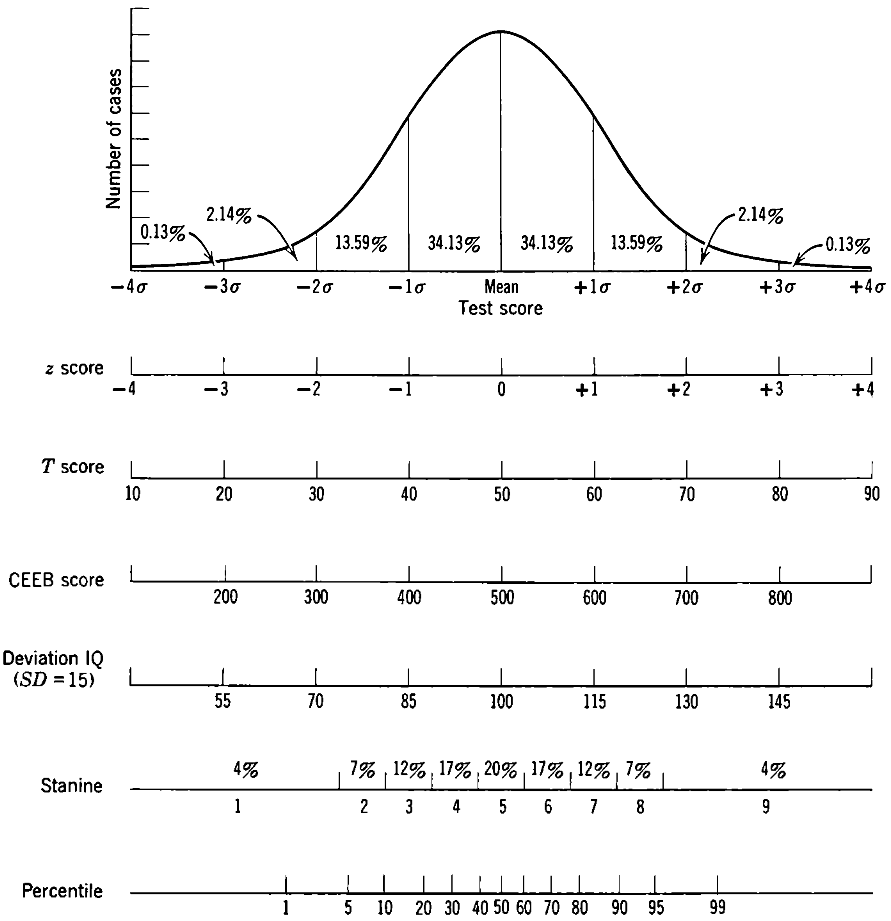


Fig. 6. Relationships among Different Types of Test Scores in a Normal Distribution.

falls at a distance of 1 *SD* above the mean and represents a standard score of +1.00. Similarly, an IQ of 132 corresponds to a standard score of +2.00, an IQ of 76 to a standard score of −1.50, and so forth. Moreover, a Stanford-Binet ratio IQ of 116 corresponds to a percentile rank of approximately 84, because in a normal curve 84 percent of the cases fall below +1.00 *SD* (Figure 4).

In Figure 6 are summarized the relationships that exist in a normal distribution among the types of scores so far discussed in this chapter. These include *z* scores, College Entrance Examination Board (CEEB) scores, Wechsler deviation IQ's (*SD* = 15), *T* scores, stanines, and percentiles. Ratio IQ's on any test will coincide with the given deviation IQ scale if they are normally distributed and have an *SD* of 15. Any other normally distributed IQ could be added to the chart, provided we know its *SD*. If the *SD* is 20, for instance, then an IQ of 120 corresponds to +1 *SD*, an IQ of 80 to −1 *SD*, and so on.

In conclusion, the exact form in which scores are reported is dictated largely by convenience, familiarity, and ease of developing norms. Standard scores in any form (including the deviation IQ) have generally replaced other types of scores because of certain advantages they offer with regard to test construction and statistical treatment of data. Most types of within-group derived scores, however, are fundamentally similar if carefully derived and properly interpreted. When certain statistical conditions are met, each of these scores can be readily translated into any of the others.

RELATIVITY OF NORMS

INTERTEST COMPARISONS. An IQ, or any other score, should always be accompanied by the name of the test on which it was obtained. Test scores cannot be properly interpreted in the abstract; they must be referred to particular tests. If the school records show that Bill Jones received an IQ of 94 and Tom Brown an IQ of 110, such IQ's cannot be accepted at face value without further information. The positions of these two students might have been reversed by exchanging the particular tests that each was given in his respective school.

Similarly, an individual's relative standing in different functions may be grossly misrepresented through lack of comparability of test norms. Let us suppose that a student has been given a verbal comprehension test and a spatial aptitude test to determine his relative standing in the two fields. If the verbal ability test was standardized on a random sample of high school students, while the spatial test was standardized on a selected group of boys attending elective shop courses, the examiner might erroneously conclude that the individual is much more able along verbal than along spatial lines, when the reverse may actually be the case.

Still another example involves longitudinal comparisons of a single individual's test performance over time. If a schoolchild's cumulative record shows IQ's of 118, 115, and 101 at the fourth, fifth, and sixth grades, the first question to ask before interpreting these changes is, "What tests did he take on these three occasions?" The apparent decline may reflect no more than the differences among the tests. In that case, he would have obtained these scores even if the three tests had been administered within a week of each other.

There are three principal reasons to account for systematic variations among the scores obtained by the same individual on different tests. First, tests may differ in *content* despite their similar labels. So-called intelligence tests provide many illustrations of this confusion. Although commonly described by the same blanket term, one of these tests may include only verbal content, another may tap predominantly spatial aptitudes, and still another may cover verbal, numerical, and spatial content in about equal proportions. Second, the *scale units* may not be comparable. As explained earlier in this chapter, if IQ's on one test have an *SD* of 12 and IQ's on another have an *SD* of 18, then an individual who received an IQ of 112 on the first test is most likely to receive an IQ of 118 on the second. Third, the composition of the *standardization samples* used in establishing norms for different tests may vary. Obviously, the same individual will appear to have performed better when compared with an inferior group than when compared with a superior group.

Lack of comparability of either test content or scale units can usually be detected by reference to the test itself or to the test manual. Differences in the respective normative samples, however, are more likely to be overlooked. Such differences probably account for many otherwise unexplained discrepancies in test results.

THE NORMATIVE SAMPLE. Any norm, however expressed, is restricted to the particular normative population from which it was derived. The test user should never lose sight of the way in which norms are established. Psychological test norms are in no sense absolute, universal, or permanent. They merely represent the test performance of the subjects constituting the standardization sample. In choosing such a sample, an effort is usually made to obtain a representative cross section of the population for which the test is designed.

In statistical terminology, a distinction is made between *sample* and *population*. The former refers to the group of individuals actually tested. The latter designates the larger, but similarly constituted, group from which the sample is drawn. For example, if we wish to establish norms of test performance for the population of 10-year-old, urban, public school boys, we might test a carefully chosen sample of 500 10-year-old boys attending public schools in several American cities. The sample would be checked with reference to geographical distribution, socioeconomic level, ethnic composition, and other relevant characteristics to ensure that it was truly representative of the defined population.

In the development and application of test norms, considerable attention should be given to the standardization sample. It is apparent that the sample on which the norms are based should be large enough to provide stable values. Another, similarly chosen sample of the same population should not yield norms that diverge appreciably from those obtained.

Norms with a large sampling error would obviously be of little value in the interpretation of test scores.

Equally important is the requirement that the sample be representative of the population under consideration. Subtle selective factors that might make the sample unrepresentative should be carefully investigated. A number of such selective factors are illustrated in institutional samples. Because such samples are usually large and readily available for testing purposes, they offer an alluring field for the accumulation of normative data. The special limitations of these samples, however, should be carefully analyzed. Testing subjects in school, for example, will yield an increasingly superior selection of cases in the successive grades, owing to the progressive dropping out of the less able pupils. Nor does such elimination affect different subgroups equally. For example, the rate of selective elimination from school is greater for boys than for girls, and it is greater in lower than in higher socioeconomic levels.

Selective factors likewise operate in other institutional samples, such as prisoners, patients in mental hospitals, or institutionalized mental retardates. Because of many special factors that determine institutionalization itself, such groups are not representative of the entire population of criminals, psychotics, or mental retardates. For example, mental retardates with physical handicaps are more likely to be institutionalized than are the physically fit. Similarly, the relative proportion of severely retarded persons will be much greater in institutional samples than in the total population.

Closely related to the question of representativeness of sample is the need for defining the specific population to which the norms apply. Obviously, one way of ensuring that a sample is representative is to restrict the population to fit the specifications of the available sample. For example, if the population is defined to include only 14-year-old schoolchildren rather than all 14-year-old children, then a school sample would be representative. Ideally, of course, the desired population should be defined in advance in terms of the objectives of the test. Then a suitable sample should be assembled. Practical obstacles in obtaining subjects, however, may make this goal unattainable. In such a case, it is far better to redefine the population more narrowly than to report norms on an ideal population which is not adequately represented by the standardization sample. In actual practice, very few tests are standardized on such broad populations as is popularly assumed. No test provides norms for the human species! And it is doubtful whether any tests give truly adequate norms for such broadly defined populations as "adult American men," "10-year-old American children," and the like. Consequently, the samples obtained by different test constructors often tend to be unrepresentative of their alleged populations and biased in different ways. Hence, the resulting norms are not comparable.

NATIONAL ANCHOR NORMS. One solution for the lack of comparability of norms is to use an anchor test to work out equivalency tables for scores on different tests. Such tables are designed to show what score in Test A is equivalent to each score in Test B. This can be done by the *equipercentile method,* in which scores are considered equivalent when they have equal percentiles in a given group. For example, if the 80th percentile in the same group corresponds to an IQ of 115 on Test A and to an IQ of 120 on Test B, then Test-A-IQ 115 is considered to be equivalent to Test-B-IQ 120. This approach has been followed to a limited extent by some test publishers, who have prepared equivalency tables for a few of their own tests (see, e.g., Lennon, 1966a).

More ambitious proposals have been made from time to time for calibrating each new test against a single anchor test, which has itself been administered to a highly representative, national normative sample (Lennon, 1966b). No single anchor test, of course, could be used in establishing norms for *all* tests, regardless of content. What is required is a battery of anchor tests, all administered to the same national sample. Each new test could then be checked against the most nearly similar anchor test in the battery.

The data gathered in Project TALENT (Flanagan et al., 1964) so far come closest to providing such an anchor battery for a high school population. Using a random sample of about 5 percent of the high schools in this country, the investigators administered a two-day battery of specially constructed aptitude, achievement, interest, and temperament tests to approximately 400,000 students in grades 9 through 12. Even with the availability of anchor data such as these, however, it must be recognized that independently developed tests can never be regarded as completely interchangeable.[4] At best, the use of national anchor norms would appreciably *reduce* the lack of comparability among tests, but it would not eliminate it.

The Project TALENT battery has been employed to calibrate several test batteries in use by the Navy and Air Force (Dailey, Shaycoft, & Orr, 1962; Shaycoft, Neyman, & Dailey, 1962). The general procedure is to administer both the Project TALENT battery and the tests to be calibrated to the same sample. Through correlational analysis, a composite of Project TALENT tests is identified that is most nearly comparable to each test to be normed. By means of the equipercentile method, tables are then prepared giving the corresponding scores on the Project TALENT composite and on the particular test. For several other batteries, data have been gathered to identify the Project TALENT composite corresponding to each test in the battery (Cooley, 1965; Cooley & Miller, 1965). These batteries include the General Aptitude Test Battery of the United States Employment Service, the Differential Aptitude Tests, and the Flanagan Aptitude Classification Tests.

---

[4] For an excellent analysis of some of the technical difficulties involved in efforts to achieve score comparability with different tests, see Angoff (1964, 1966, 1971a).

Of particular interest is The Anchor Test Study conducted by the Educational Testing Service under the auspices of the U.S. Office of Education (Jaeger, 1973). This study represents a systematic effort to provide comparable and truly representative national norms for the seven most widely used reading achievement tests for elementary schoolchildren. Through an unusually well-controlled experimental design, over 300,000 fourth-, fifth-, and sixth-grade schoolchildren were examined in 50 states. The anchor test consisted of the reading comprehension and vocabulary subtests of the Metropolitan Achievement Test, for which new norms were established in one phase of the project. In the equating phase of the study, each child took the reading comprehension and vocabulary subtests from two of the seven batteries, each battery being paired in turn with every other battery. Some groups took parallel forms of the two subtests from the same battery. In still other groups, all the pairings were duplicated in reverse sequence, in order to control for order of administration. From statistical analyses of all these data, score equivalency tables for the seven tests were prepared by the equipercentile method. A manual for interpreting scores is provided for use by school systems and other interested persons (Loret, Seder, Bianchini, & Vale, 1974).

SPECIFIC NORMS. Another approach to the nonequivalence of existing norms—and probably a more realistic one for most tests—is to standardize tests on more narrowly defined populations, so chosen as to suit the specific purposes of each test. In such cases, the limits of the normative population should be clearly reported with the norms. Thus, the norms might be said to apply to "employed clerical workers in large business organizations" or to "first-year engineering students." For many testing purposes, highly specific norms are desirable. Even when representative norms are available for a broadly defined population, it is often helpful to have separately reported *subgroup norms*. This is true whenever recognizable subgroups yield appreciably different scores on a particular test. The subgroups may be formed with respect to age, grade, type of curriculum, sex, geographical region, urban or rural environment, socioeconomic level, and many other factors. The use to be made of the test determines the type of differentiation that is most relevant, as well as whether general or specific norms are more appropriate.

Mention should also be made of *local norms*, often developed by the test users themselves within a particular setting. The groups employed in deriving such norms are even more narrowly defined than the subgroups considered above. Thus, an employer may accumulate norms on applicants for a given type of job within his company. A college admissions office may develop norms on its own student population. Or a single elementary school may evaluate the performance of individual pupils in terms of its own score distribution. These local norms are more appropriate than broad national norms for many testing purposes, such as the prediction of subsequent job performance or college achievement, the comparison of a child's relative achievement in different subjects, or the measurement of an individual's progress over time.

FIXED REFERENCE GROUP. Although most derived scores are computed in such a way as to provide an immediate normative interpretation of test performance, there are some notable exceptions. One type of nonnormative scale utilizes a fixed reference group in order to ensure *comparability and continuity* of scores, without providing normative evaluation of performance. With such a scale, normative interpretation requires reference to independently collected norms from a suitable population. Local or other specific norms are often used for this purpose.

One of the clearest examples of scaling in terms of a fixed reference group is provided by the score scale of the College Board Scholastic Aptitude Test (Angoff, 1962, 1971b). Between 1926 (when this test was first administered) and 1941, SAT scores were expressed on a normative scale, in terms of the mean and *SD* of the candidates taking the test at each administration. As the number and variety of College Board member colleges increased and the composition of the candidate population changed, it was concluded that scale continuity should be maintained. Otherwise, an individual's score would depend on the characteristics of the group tested during a particular year. An even more urgent reason for scale continuity stemmed from the observation that students taking the SAT at certain times of the year performed more poorly than those taking it at other times, owing to the differential operation of selective factors. After 1941, therefore, all SAT scores were expressed in terms of the mean and *SD* of the approximately 11,000 candidates who took the test in 1941. These candidates constitute the fixed reference group employed in scaling all subsequent forms of the test. Thus, a score of 500 on any form of the SAT corresponds to the mean of the 1941 sample; a score of 600 falls 1 *SD* above that mean, and so forth.

To permit translation of raw scores on any form of the SAT into these fixed-reference-group scores, a short anchor test (or set of common items) is included in each form. Each new form is thereby linked to one or two earlier forms, which in turn are linked with other forms by a chain of items extending back to the 1941 form. These nonnormative SAT scores can then be interpreted by comparison with any appropriate distribution of scores, such as that of a particular college, a type of college, a region, etc. These specific norms are more useful in making college admission decisions than would be annual norms based on the entire candidate population. Any changes in the candidate population over time, moreover, can be detected only with a fixed-score scale. It will be noted that the principal difference between the fixed-reference-group scales under consideration and the previously discussed scales based on national anchor norms is that the latter require the choice of a single group that is broadly representative and appropriate for normative purposes. Apart from the practical difficulties in obtaining such a group and the need to update the norms, it is likely that for many testing purposes such broad norms are not required.

Scales built from a fixed reference group are analogous in one respect to scales employed in physical measurement. In this connection, Angoff (1962, pp. 32–33) writes:

> There is hardly a person here who knows the precise original definition of the length of the foot used in the measurement of height or distance, or which king it was whose foot was originallv agreed upon as the standard; on the other hand, there is no one here who does not know how to evaluate lengths and distances in terms of this unit. Our ignorance of the precise original meaning or derivation of the foot does not lessen its usefulness to us in any way. Its usefulness derives from the fact that it remains the same over time and allows us to familiarize ourselves with it. Needless to say, precisely the same considerations apply to other units of measurement—the inch, the mile, the degree of Fahrenheit, and so on. In the field of psychological measurement it is similarly reasonable to say that the original definition of the scale is or should be of no consequence. What is of consequence is the maintenance of a constant scale—which, in the case of a multiple-form testing program, is achieved bv rigorous form-to-form equating—and the provision of supplementary normative data to aid in interpretation and in the formation of specific decisions, data which would be revised from time to time as conditions warrant.

COMPUTER UTILIZATION IN THE INTERPRETATION OF TEST SCORES

Computers have already made a significant impact upon every phase of testing, from test construction to administration, scoring, reporting, and interpretation. The obvious uses of computers—and those developed earliest—represent simply an unprecedented increase in the speed with which traditional data analyses and scoring processes can be carried out. Far more important, however, are the adoption of new procedures and the exploration of new approaches to psychological testing which would have been impossible without the flexibility, speed, and data-processing capabilities of computers. As Baker (1971, p. 227) succinctly puts it, computer capabilities should serve "to free one's thinking from the constraints of the past."

Various testing innovations resulting from computer utilization will be discussed under appropriate topics throughout the book. In the present connection, we shall examine some applications of computers in the interpretation of test scores. At the simplest level, most current tests, and especially those designed for group administration, are now adapted for *computer scoring* (Baker, 1971). Several test publishers, as well as independent test-scoring organizations, are equipped to provide such scoring services to test users. Although separate answer sheets are commonly used for this purpose, optical scanning equipment available at some scoring centers permits the reading of responses directly from test booklets. Many innovative possibilities, such as diagnostic scoring and path analysis (recording a student's progress at various stages of learning) have barely been explored.

At a somewhat more complex level, certain tests now provide facilities for *computer interpretation* of test scores. In such cases, the computer program associates prepared verbal statements with particular patterns of test responses. This approach has been pursued with both personality and aptitude tests. For example, with the Minnesota Multiphasic Personality Inventory (MMPI), to be discussed in Chapter 17, test users may obtain computer printouts of diagnostic and interpretive statements about the subject's personality tendencies and emotional condition, together with the numerical scores. Similarly, the Differential Aptitude Tests (see Ch. 13) provide a Career Planning Report, which includes a profile of scores on the separate subtests as well as an interpretive computer printout. The latter contains verbal statements that combine the test data with information on interests and goals given by the student on a Career Planning Questionnaire. These statements are typical of what a counselor would say to the student in going over his test results in an individual conference (Super, 1973).

Individualized interpretation of test scores at a still more complex level is illustrated by *interactive computer systems,* in which the individual is in direct contact with the computer by means of response stations and in effect engages in a dialogue with the computer (J. A. Harris, 1973; Holtzman, 1970; M. R. Katz, 1974; Super, 1970). This technique has been investigated with regard to educational and vocational planning and decision making. In such a situation, test scores are usually incorporated in the computer data base, together with other information provided by the student or client. Essentially, the computer combines all the available information about the individual with stored data about educational programs and occupations; and it utilizes all relevant facts and relations in answering the individual's questions and aiding him in reaching decisions. Examples of such interactive computer systems, in various stages of operational development, include IBM's Education and Career Exploration System (ECES) and ETS's System for Interactive Guidance Information (SIGI). Preliminary field trials show good acceptance of these systems by high school students and their parents (Harris, 1973).

Test results also represent an integral part of the data utilized in *computer-assisted instruction* (CAI).[5] In order to present instructional material appropriate to each student's current level of attainment, the computer must repeatedly score and evaluate the student's responses to preceding material. On the basis of his response history, the student may be routed to more advanced material, or to further practice at the present level, or to a remedial branch whereby he receives instruction in more elementary prerequisite material. Diagnostic analysis of errors may lead to an instructional program designed to correct the specific learning difficulties identified in individual cases.

A less costly and operationally more feasible variant of computer utilization in learning is *computer-managed instruction* (CMI—see Hambleton, 1974). In such systems, the learner does not interact directly with the computer. The role of the computer is to assist the teacher in carrying out a plan of individualized instruction, which may use self-instruction packages or more conventional types of instruction. A major contribution of the computer is to process the rather formidable mass of data accumulated daily regarding the performance of each student, in a classroom where each may be involved in a different activity, and to utilize these data in prescribing the next instructional step for each student. Examples of this application of computers are provided by the University of Pittsburgh's IPI (Individually Prescribed Instruction—see Cooley & Glaser, 1969; Glaser, 1968) and Project PLAN (Planning for Learning in Accordance with Needs) developed by the American Institutes for Research (Flanagan, 1971; Flanagan, Shanner, Brudner, & Marker, 1975). Project PLAN includes a program of self-knowledge, individual development, and occupational planning, as well as instruction in elementary and high school subjects.

CRITERION-REFERENCED TESTING

NATURE AND USES. An approach to testing that has aroused a surge of activity, particularly in education, is generally designated as "criterion-referenced testing." First proposed by Glaser (1963), this term is still used somewhat loosely and its definition varies among different writers. Moreover, several alternative terms are in common use, such as content-, domain-, and objective-referenced. These terms are sometimes employed as synonyms for criterion-referenced and sometimes with slightly different connotations. "Criterion-referenced," however, seems to have gained ascendancy, although it is not the most appropriate term.

---

[5] For a description of a widely used CAI system for teaching reading to first-, second-, and third-grade children, see R. C. Atkinson (1974).

Typically, criterion-referenced testing uses as its interpretive frame of reference a specified *content* domain rather than a specified population of *persons*. In this respect, it has been contrasted with the usual norm-referenced testing, in which an individual's score is interpreted by comparing it with the scores obtained by others on the same test. In criterion-referenced testing, for example, an examinee's test performance may be reported in terms of the specific kinds of arithmetic operations he has mastered, the estimated size of his vocabulary, the difficulty level of reading matter he can comprehend (from comic books to literary classics), or the chances of his achieving a designated performance level on an external criterion (educational or vocational).

Thus far, criterion-referenced testing has found its major applications in several recent innovations in education. Prominent among these are computer-assisted, computer-managed, and other individualized, self-paced instructional systems. In all these systems, testing is closely integrated with instruction, being introduced before, during, and after completion of each instructional unit to check on prerequisite skills, diagnose possible learning difficulties, and prescribe subsequent instructional procedures. The previously cited Project PLAN and IPI are examples of such programs.

From another angle, criterion-referenced tests are useful in broad surveys of educational accomplishment, such as the National Assessment of Educational Progress (Womer, 1970), and in meeting demands for educational accountability (Gronlund, 1974). From still another angle, testing for the attainment of minimum requirements, as in qualifying for a driver's license or a pilot's license, illustrates criterion-referenced testing. Finally, familiarity with the concepts of criterion-referenced testing can contribute to the improvement of the traditional, informal tests prepared by teachers for classroom use. Gronlund (1973) provides a helpful guide for this purpose, as well as a simple and well-balanced introduction to criterion-referenced testing. A brief but excellent discussion of the chief limitations of criterion-referenced tests is given by Ebel (1972b).

CONTENT MEANING. The major distinguishing feature of criterion-referenced testing (however defined and whether designated by this term or by one of its synonyms) is its interpretation of test performance in terms of content meaning. The focus is clearly on *what* the person can do and what he knows, not on how he compares with others. A fundamental requirement in constructing this type of test is a clearly defined domain of knowledge or skills to be assessed by the test. If scores on such a test are to have communicable meaning, the content domain to be sampled must be widely recognized as important. The selected domain must then be subdivided into small units defined in performance terms. In an educational context, these units correspond to behaviorally defined instructional objectives, such as "multiplies three-digit by two-digit numbers" or "identifies the misspelled word in which the final *e* is retained when adding *-ing*." In the programs prepared for individualized instruction, these objectives run to several hundred for a single school subject. After the instructional objectives have been formulated, items are prepared to sample each objective. This procedure is admittedly difficult and time consuming. Without such careful specification and control of content, however, the results of criterion-referenced testing could degenerate into an idiosyncratic and uninterpretable jumble.

When strictly applied, criterion-referenced testing is best adapted for testing basic skills (as in reading and arithmetic) at elementary levels. In these areas, instructional objectives can also be arranged in an ordinal hierarchy, the acquisition of more elementary skills being prerequisite to the acquisition of higher-level skills.[6] It is impracticable and probably undesirable, however, to formulate highly specific objectives for advanced levels of knowledge in less highly structured subjects. At these levels, both the content and sequence of learning are likely to be much more flexible.

On the other hand, in its emphasis on content meaning in the interpretation of test scores, criterion-referenced testing may exert a salutary effect on testing in general. The interpretation of intelligence test scores, for example, would benefit from this approach. To describe a child's intelligence test performance in terms of the specific intellectual skills and knowledge it represents might help to counteract the confusions and misconceptions that have become attached to the IQ. When stated in these general terms, however, the criterion-referenced approach is equivalent to interpreting test scores in the light of the demonstrated validity of the particular test, rather than in terms of vague underlying entities. Such an interpretation can certainly be combined with norm-referenced scores.

MASTERY TESTING. A second major feature almost always found in criterion-referenced testing is the procedure of testing for mastery. Essentially, this procedure yields an all-or-none score, indicating that the individual has or has not attained the preestablished level of mastery. When basic skills are tested, nearly complete mastery is generally expected (e.g., 80–85% correct items). A three-way distinction may also be employed, including mastery, nonmastery, and an intermediate, doubtful, or "review" interval.

---

[6] Ideally, such tests follow the simplex model of a Guttman scale (see Popham & Husek, 1969), as do the Piagetian ordinal scales discussed earlier in this chapter.

In connection with individualized instruction, some educators have argued that, given enough time and suitable instructional methods, nearly everyone can achieve complete mastery of the chosen instructional objectives. Individual differences would thus be manifested in learning time rather than in final achievement as in traditional educational testing (Bloom, 1968; J. B. Carroll, 1963, 1970; Cooley & Glaser, 1969; Gagné, 1965). It follows that in mastery testing, individual differences in performance are of little or no interest. Hence as generally constructed, criterion-referenced tests minimize individual differences. For example, they include items passed or failed by all or nearly all examinees, although such items are usually excluded from norm-referenced tests.[7] Mastery testing is regularly employed in the previously cited programs for individualized instruction. It is also characteristic of published criterion-referenced tests for basic skills, suitable for elementary school. Examples of such tests include the Prescriptive Reading Inventory and Prescriptive Mathematics Inventory (California Test Bureau), The Skills Monitoring System in Reading and in Study Skills (Harcourt Brace Jovanovich), and Diagnosis: An Instructional Aid Series in Reading and in Mathematics (Science Research Associates).

Beyond basic skills, mastery testing is inapplicable or insufficient. In more advanced and less structured subjects, achievement is open-ended. The individual may progress almost without limit in such functions as understanding, critical thinking, appreciation, and originality. Moreover, content coverage may proceed in many different directions, depending upon the individual's abilities, interests, and goals, as well as local instructional facilities. Under these conditions, complete mastery is unrealistic and unnecessary. Hence norm-referenced evaluation is generally employed in such cases to assess degree of attainment. Some published tests are so constructed as to permit both norm-referenced and criterion-referenced applications. An example is the 1973 Edition of the Stanford Achievement Test. While providing appropriate norms at each level, this battery meets three important requirements of criterion-referenced tests: specification of detailed instructional objectives, adequate coverage of each objective with appropriate items, and wide range of item difficulty.

It should be noted that criterion-referenced testing is neither as new nor as clearly divorced from norm-referenced testing as some of its proponents imply. Evaluating an individual's test performance in absolute terms, such as by letter grades or percentage of correct items, is certainly much older than normative interpretations. More precise attempts to describe test performance in terms of content meaning also antedate the introduction of the term "criterion-referenced testing" (Ebel, 1962; Flanagan, 1962—see also Anastasi, 1968, pp. 69–70). Other examples may be found in early product scales for assessing the quality of handwriting, compositions, or drawings by matching the individual's work sample against a set of standard specimens. Ebel (1972b) observes, furthermore, that the concept of mastery in education—in the sense of all-or-none learning of specific units—achieved considerable popularity in the 1920s and 1930s and was later abandoned.

---

[7] As a result of this reduction in variability, the usual methods for finding reliability and validity are inapplicable to most criterion-referenced tests. Further discussion of these points will be found in Chapters 5, 6, and 8.

A normative framework is implicit in all testing, regardless of how scores are expressed (Angoff, 1974). The very choice of content or skills to be measured is influenced by the examiner's knowledge of what can be expected from human organisms at a particular developmental or instructional stage. Such a choice presupposes information about what other persons have done in similar situations. Moreover, by imposing uniform cutoff scores on an ability continuum, mastery testing does not thereby eliminate individual differences. To describe an individual's level of reading comprehension as "the ability to understand the content of the *New York Times*" still leaves room for a wide range of individual differences in degree of understanding.

EXPECTANCY TABLES. Test scores may also be interpreted in terms of expected criterion performance, as in a training program or on a job. This usage of the term "criterion" follows standard psychometric practice, as when a test is said to be validated against a particular criterion (see Ch. 2). Strictly speaking, the term "criterion-referenced testing" should refer to this type of performance interpretation, while the other approaches discussed in this section can be more precisely described as content-referenced. This terminology, in fact, is used in the APA test *Standards* (1974).

An expectancy table gives the probability of different criterion outcomes for persons who obtain each test score. For example, if a student obtains a score of 530 on the CEEB Scholastic Aptitude Test, what are the chances that his freshman grade-point average in a specific college will fall in the A, B, C, D, or F category? This type of information can be obtained by examining the bivariate distribution of predictor scores (SAT) plotted against criterion status (freshman grade-point average). If the number of cases in each cell of such a bivariate distribution is changed to a percentage, the result is an expectancy table, such as the one illustrated in Table 6. The data for this table were obtained from 171 high school boys enrolled in courses in American history. The predictor was the Verbal Reasoning test of the Differential Aptitude Tests, administered early in the course. The criterion was end-of-course grades. The correlation between test scores and criterion was .66.

TABLE 6

Expectancy Table Showing Relation between DAT Verbal Reasoning Test and Course Grades in American History for 171 Boys in Grade 11

(Adapted from Fifth Edition Manual for the Differential Aptitude Tests, Forms S and T, p. 114. Reproduced by permission. Copyright © 1973, 1974 by The Psychological Corporation, New York, N.Y. All rights reserved.)

Test Score	Number of Cases	Percentage Receiving Each Criterion Grade			
		Below 70	70–79	80–89	90 & above
40 & above	46		15	22	63
30–39	36	6	39	39	17
20–29	43	12	63	21	5
Below 20	46	30	52	17	

The first column of Table 6 shows the test scores, divided into four class intervals; the number of students whose scores fall into each interval is given in the second column. The remaining entries in each row of the table indicate the percentage of cases within each test-score interval who received each grade at the end of the course. Thus, of the 46 students with scores of 40 or above on the Verbal Reasoning test, 15 percent received grades of 70–79, 22 percent grades of 80–89, and 63 percent grades of 90 or above. At the other extreme, of the 46 students scoring below 20 on the test, 30 percent received grades below 70, 52 percent between 70 and 79, and 17 percent between 80 and 89. Within the limitations of the available data, these percentages represent the best estimates of the probability that an individual will receive a given criterion grade. For example, if a new student receives a test score of 34 (i.e., in the 30–39 interval), we would conclude that the probability of his obtaining a grade of 90 or above is 17 out of 100; the probability of his obtaining a grade between 80 and 89 is 39 out of 100, and so on.

In many practical situations, criteria can be dichotomized into "success" and "failure" in a job, course of study, or other undertaking. Under these conditions, an *expectancy chart* can be prepared, showing the probability of success or failure corresponding to each score interval. Figure 7 is an example of such an expectancy chart. Based on a pilot selection battery developed by the Air Force, this expectancy chart shows

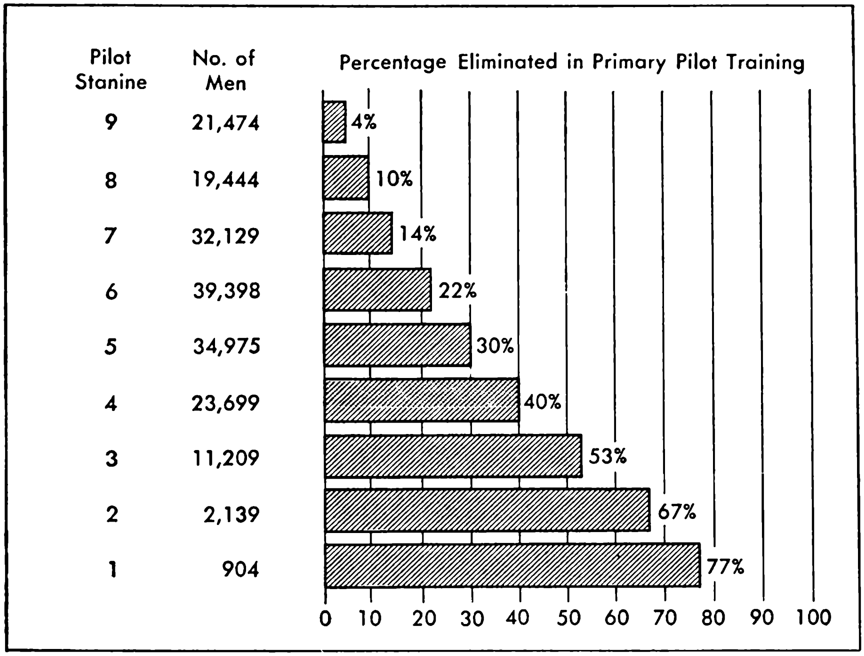


FIG. 7. Expectancy Chart Showing Relation between Performance on Pilot Selection Battery and Elimination from Primary Flight Training.

(From Flanagan, 1947, p. 58.)

the percentage of men scoring within each stanine on the battery who failed to complete primary flight training. It can be seen that 77 percent of the men receiving a stanine of 1 were eliminated in the course of training, while only 4 percent of those at stanine 9 failed to complete the training satisfactorily. Between these extremes, the percentage of failures decreases consistently over the successive stanines. On the basis of this expectancy chart, it could be predicted, for example, that approximately 40 percent of pilot cadets who obtain a stanine score of 4 will fail and approximately 60 percent will satisfactorily complete primary flight training. Similar statements regarding the probability of success and failure could be made about individuals who receive each stanine. Thus, an individual with a stanine of 4 has a 60:40 or 3:2 chance of completing primary flight training. Besides providing a criterion-referenced interpretation of test scores, it can be seen that both expectancy tables and expectancy charts give a general idea of the validity of a test in predicting a given criterion.

CHAPTER 5

Reliability

RELIABILITY refers to the consistency of scores obtained by the same persons when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions. This concept of reliability underlies the computation of the *error of measurement* of a single score, whereby we can predict the range of fluctuation likely to occur in a single individual's score as a result of irrelevant, chance factors.

The concept of test reliability has been used to cover several aspects of score consistency. In its broadest sense, test reliability indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. To put it in more technical terms, measures of test reliability make it possible to estimate what proportion of the total variance of test scores is *error variance.* The crux of the matter, however, lies in the definition of error variance. Factors that might be considered error variance for one purpose would be classified under true variance for another. For example, if we are interested in measuring fluctuations of mood, then the day-by-day changes in scores on a test of cheerfulness-depression would be relevant to the purpose of the test and would hence be part of the true variance of the scores. If, on the other hand, the test is designed to measure more permanent personality characteristics, the same daily fluctuations would fall under the heading of error variance.

Essentially, any condition that is irrelevant to the purpose of the test represents error variance. Thus, when the examiner tries to maintain uniform testing conditions by controlling the testing environment, instructions, time limits, rapport, and other similar factors, he is reducing error variance and making the test scores more reliable. Despite optimum testing conditions, however, no test is a perfectly reliable instrument. Hence, every test should be accompanied by a statement of its reliability. Such a measure of reliability characterizes the test when administered under standard conditions and given to subjects similar to those constituting the normative sample. The characteristics of this sample should therefore be specified, together with the type of reliability that was measured.

There could, of course, be as many varieties of test reliability as there are conditions affecting test scores, since any such conditions might be irrelevant for a certain purpose and would thus be classified as error variance. The types of reliability computed in actual practice, however, are relatively few. In this chapter, the principal techniques for measuring the reliability of test scores will be examined, together with the sources of error variance identified by each. Since all types of reliability are concerned with the degree of consistency or agreement between two independently derived sets of scores, they can all be expressed in terms of a *correlation coefficient.* Accordingly, the next section will consider some of the basic characteristics of correlation coefficients, in order to clarify their use and interpretation. More technical discussion of correlation, as well as more detailed specifications of computing procedures, can be found in any elementary textbook of educational or psychological statistics, such as Guilford and Fruchter (1973).

THE CORRELATION COEFFICIENT

MEANING OF CORRELATION. Essentially, a correlation coefficient (r) expresses the degree of correspondence, or *relationship,* between two sets of scores. Thus, if the top-scoring individual in variable 1 also obtains the top score in variable 2, the second-best individual in variable 1 is second best in variable 2, and so on down to the poorest individual in the group, then there would be a perfect correlation between variables 1 and 2. Such a correlation would have a value of +1.00.

A hypothetical illustration of a perfect positive correlation is shown in Figure 8. This figure presents a scatter diagram, or bivariate distribution. Each tally mark in this diagram indicates the score of one individual in both variable 1 (horizontal axis) and variable 2 (vertical axis). It will be noted that all of the 100 cases in the group are distributed along the diagonal running from the lower left- to the upper right-hand corner of the diagram. Such a distribution indicates a perfect positive correlation (+1.00), since it shows that each individual occupies the same relative position in both variables. The closer the bivariate distribution of scores approaches this diagonal, the higher will be the positive correlation.

Figure 9 illustrates a perfect negative correlation (−1.00). In this case, there is a complete reversal of scores from one variable to the other. The best individual in variable 1 is the poorest in variable 2 and vice versa, this reversal being consistently maintained throughout the distribution. It will be noted that, in this scatter diagram, all individuals fall on the diagonal extending from the upper left- to the lower right-hand corner. This diagonal runs in the reverse direction from that in Figure 8.

A zero correlation indicates complete absence of relationship, such as

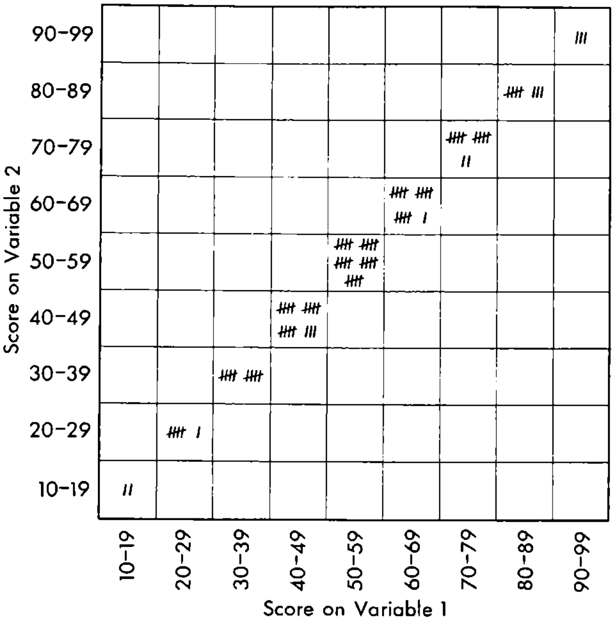


FIG. 8. Bivariate Distribution for a Hypothetical Correlation of +1.00.

might occur by chance. If each individual's name were pulled at random out of a hat to determine his position in variable 1, and if the process were repeated for variable 2, a zero or near-zero correlation would result. Under these conditions, it would be impossible to predict an individual's relative standing in variable 2 from a knowledge of his score in variable 1. The top-scoring subject in variable 1 might score high, low, or average in variable 2. Some individuals might by chance score above average in both variables, or below average in both; others might fall above average in one variable and below in the other; still others might be above the average in one and at the average in the second, and so forth. There would be no regularity in the relationship from one individual to another.

The coefficients found in actual practice generally fall between these extremes, having some value higher than zero but lower than 1.00. Correlations between measures of abilities are nearly always positive, although frequently low. When a negative correlation is obtained between two such variables, it usually results from the way in which the scores are expressed. For example, if time scores are correlated with amount scores, a negative correlation will probably result. Thus, if each subject's score on an arithmetic computation test is recorded as the number of seconds required to complete all items, while his score on an arithmetic reasoning test represents the number of problems correctly solved, a negative correlation can be expected. In such a case, the poorest (i.e., slowest) individ-

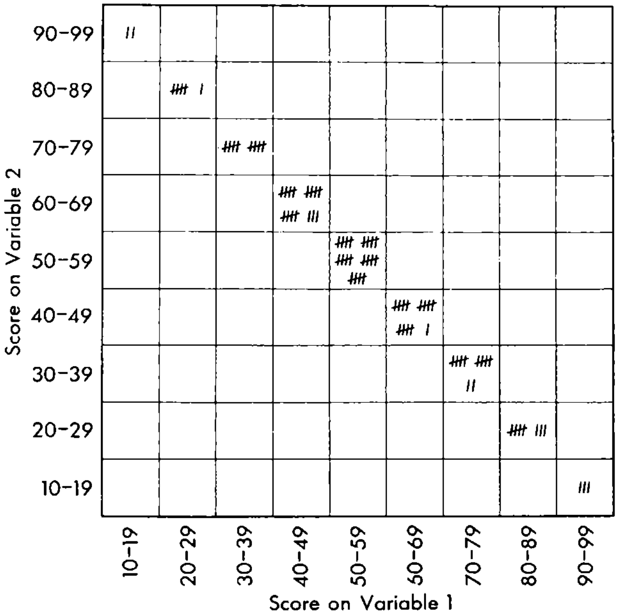


FIG. 9. Bivariate Distribution for a Hypothetical Correlation of −1.00.

ual will have the numerically highest score on the first test, while the best individual will have the highest score on the second.

Correlation coefficients may be computed in various ways, depending on the nature of the data. The most common is the *Pearson Product-Moment Correlation Coefficient*. This correlation coefficient takes into account not only the person's position in the group, but also the amount of his deviation above or below the group mean. It will be recalled that when each individual's standing is expressed in terms of standard scores, persons falling above the average receive positive standard scores, while those below the average receive negative scores. Thus, an individual who is superior in both variables to be correlated would have two positive standard scores; one inferior in both would have two negative standard scores. If, now, we multiply each individual's standard score in variable 1 by his standard score in variable 2, all of these products will be positive, provided that each individual falls on the same side of the mean on both variables. The Pearson correlation coefficient is simply the mean of these products. It will have a high positive value when corresponding standard scores are of equal sign and of approximately equal amount in the two variables. When subjects above the average in one variable are below the average in the other, the corresponding cross-products will be negative. If the sum of the cross-products is negative, the correlation will be negative. When some products are positive and some negative, the correlation will be close to zero.

In actual practice, it is not necessary to convert each raw score to a standard score before finding the cross-products, since this conversion can be made once for all after the cross-products have been added. There are many shortcuts for computing the Pearson correlation coefficient. The method demonstrated in Table 7 is not the quickest, but it illustrates the meaning of the correlation coefficient more clearly than other methods that utilize computational shortcuts. Table 7 shows the computation of a Pearson r between the arithmetic and reading scores of 10 children. Next to each child's name are his scores in the arithmetic test (X) and the reading test (Y). The sums and means of the 10 scores are given under the respective columns. The third column shows the deviation (x) of each arithmetic score from the arithmetic mean; and the fourth column, the deviation (y) of each reading score from the reading mean. These deviations are squared in the next two columns, and the sums of the squares are used in computing the standard deviations of the arithmetic and reading scores by the method described in Chapter 4. Rather than dividing each x and y by its corresponding σ to find standard scores, we

Table 7
Computation of Pearson Product-Moment Correlation Coefficient

Pupil	Arithmetic X	Reading Y	x	y	x^2	y^2	xy
Bill	41	17	+1	−4	1	16	− 4
Carol	38	28	−2	+7	4	49	−14
Geoffrey	48	22	+8	+1	64	1	8
Ann	32	16	−8	−5	64	25	40
Bob	34	18	−6	−3	36	9	18
Jane	36	15	−4	−6	16	36	24
Ellen	41	24	+1	+3	1	9	3
Ruth	43	20	+3	−1	9	1	− 3
Dick	47	23	+7	+2	49	4	14
Mary	40	27	0	+6	0	36	0
Σ	400	210	0	0	244	186	86
M	40	21					

$$\sigma_x = \sqrt{\frac{244}{10}} = \sqrt{24.40} = 4.94 \qquad \sigma_y = \sqrt{\frac{186}{10}} = \sqrt{18.60} = 4.31$$

$$r_{xy} = \frac{\Sigma xy}{N\,\sigma_x\,\sigma_y} = \frac{86}{(10)\,(4.94)\,(4.31)} = \frac{86}{212.91} = .40$$

perform this division only once at the end, as shown in the correlation formula in Table 7. The cross-products in the last column (xy) have been found by multiplying the corresponding deviations in the x and y columns. To compute the correlation (r), the sum of these cross-products is divided by the number of cases (N) and by the product of the two standard deviations ($\sigma_x \sigma_y$).

STATISTICAL SIGNIFICANCE. The correlation of .40 found in Table 7 indicates a moderate degree of positive relationship between the arithmetic and reading scores. There is some tendency for those children doing well in arithmetic also to perform well on the reading test and vice versa, although the relation is not close. If we are concerned only with the performance of these 10 children, we can accept this correlation as an adequate description of the degree of relation existing between the two variables in this group. In psychological research, however, we are usually interested in generalizing beyond the particular *sample* of individuals tested to the larger *population* which they represent. For example, we might want to know whether arithmetic and reading ability are correlated among American schoolchildren of the same age as those we tested. Obviously, the 10 cases actually examined would constitute a very inadequate sample of such a population. Another comparable sample of the same size might yield a much lower or a much higher correlation.

There are statistical procedures for estimating the probable fluctuation to be expected from sample to sample in the size of correlations, means, standard deviations, and any other group measures. The question usually asked about correlations, however, is simply whether the correlation is significantly greater than zero. In other words, if the correlation in the population is zero, could a correlation as high as that obtained in our sample have resulted from sampling error alone? When we say that a correlation is "significant at the 1 percent (.01) level," we mean the chances are no greater than one out of 100 that the population correlation is zero. Hence, we conclude that the two variables are truly correlated. Significance levels refer to the risk of error we are willing to take in drawing conclusions from our data. If a correlation is said to be significant at the .05 level, the probability of error is 5 out of 100. Most psychological research applies either the .01 or the .05 levels, although other significance levels may be employed for special reasons.

The correlation of .40 found in Table 7 fails to reach significance even at the .05 level. As might have been anticipated, with only 10 cases it is difficult to establish a general relationship conclusively. With this size of sample, the smallest correlation significant at the .05 level is .63. Any correlation below that value simply leaves unanswered the question of whether the two variables are correlated in the population from which the sample was drawn.

The minimum correlations significant at the .01 and .05 levels for groups of different sizes can be found by consulting tables of the significance of correlations in any statistics textbook. For interpretive purposes in this book, however, only an understanding of the general concept is required. Parenthetically, it might be added that significance levels can be interpreted in a similar way when applied to other statistical measures. For example, to say that the difference between two means is significant at the .01 level indicates that we can conclude, with only one chance out of 100 of being wrong, that a difference in the obtained direction would be found if we tested the whole population from which our samples were drawn. For instance, if in the sample tested the boys had obtained a significantly higher mean than the girls on a mechanical comprehension test, we could conclude that the boys would also excel in the total population.

THE RELIABILITY COEFFICIENT. Correlation coefficients have many uses in the analysis of psychological data. The measurement of test reliability represents one application of such coefficients. An example of a reliability coefficient, computed by the Pearson Product-Moment method, is to be found in Figure 10. In this case, the scores of 104 persons on two equivalent forms of a Word Fluency test[1] were correlated. In one form, the subjects were given five minutes to write as many words as they could that began with a given letter. The second form was identical, except that a different letter was employed. The two letters were chosen by the test authors as being approximately equal in difficulty for this purpose.

The correlation between the number of words written in the two forms of this test was found to be .72. This correlation is high and significant at the .01 level. With 104 cases, any correlation of .25 or higher is significant at this level. Nevertheless, the obtained correlation is somewhat lower than is desirable for reliability coefficients, which usually fall in the .80's or .90's. An examination of the scatter diagram in Figure 10 shows a typical bivariate distribution of scores corresponding to a high positive correlation. It will be noted that the tallies cluster close to the diagonal extending from the lower left- to the upper right-hand corner; the trend is definitely in this direction, although there is a certain amount of scatter of individual entries. In the following section, the use of the correlation coefficient in computing different measures of test reliability will be considered.

[1] One of the subtests of the SRA Tests of Primary Mental Abilities for Ages 11 to 17. The data were obtained in an investigation by Anastasi and Drake (1954).

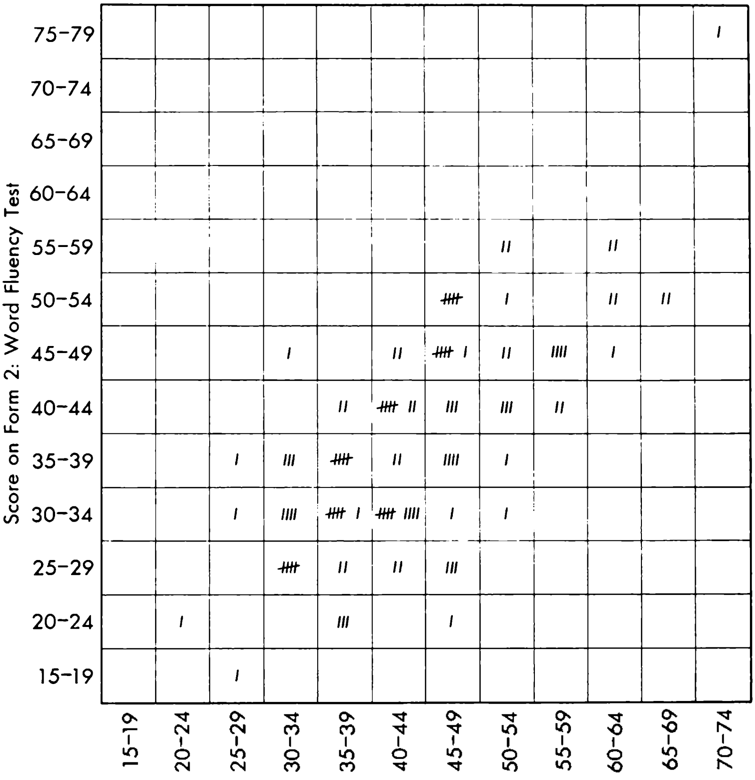


FIG. 10. A Reliability Coefficient of .72.
(Data from Anastasi & Drake, 1954.)

TYPES OF RELIABILITY

TEST-RETEST RELIABILITY. The most obvious method for finding the reliability of test scores is by repeating the identical test on a second occasion. The reliability coefficient (r_{11}) in this case is simply the correlation between the scores obtained by the same persons on the two administrations of the test. The error variance corresponds to the random fluctuations of performance from one test session to the other. These variations may result in part from uncontrolled testing conditions, such as extreme changes in weather, sudden noises and other distractions, or a broken pencil point. To some extent, however, they arise from changes in the condition of the subject himself, as illustrated by illness, fatigue, emotional strain, worry, recent experiences of a pleasant or unpleasant nature, and the like. Retest reliability shows the extent to which scores on a test can be generalized over different occasions; the higher the reliability, the less susceptible the scores are to the random daily changes in the condition of the subject or of the testing environment.

When retest reliability is reported in a test manual, the interval over which it was measured should always be specified. Since retest correlations decrease progressively as this interval lengthens, there is not one but an infinite number of retest reliability coefficients for any test. It is also desirable to give some indication of relevant intervening experiences of the subjects on whom reliability was measured, such as educational or job experiences, counseling, psychotherapy, and so forth.

Apart from the desirability of reporting length of interval, what considerations should guide the choice of interval? Illustrations could readily be cited of tests showing high reliability over periods of a few days or weeks, but whose scores reveal an almost complete lack of correspondence when the interval is extended to as long as ten or fifteen years. Many preschool intelligence tests, for example, yield moderately stable measures within the preschool period, but are virtually useless as predictors of late childhood or adult IQ's. In actual practice, however, a simple distinction can usually be made. Short-range, random fluctuations that occur during intervals ranging from a few hours to a few months are generally included under the error variance of the test score. Thus, in checking this type of test reliability, an effort is made to keep the interval short. In testing young children, the period should be even shorter than for older persons, since at early ages progressive developmental changes are discernible over a period of a month or even less. For any type of person, the interval between retests should rarely exceed six months.

Any additional changes in the relative test performance of individuals that occur over longer periods of time are apt to be cumulative and progressive rather than entirely random. Moreover, they are likely to characterize a broader area of behavior than that covered by the test performance itself. Thus, one's general level of scholastic aptitude, mechanical comprehension, or artistic judgment may have altered appreciably over a ten-year period, owing to unusual intervening experiences. The individual's status may have either risen or dropped appreciably in relation to others of his own age, because of circumstances peculiar to his own home, school, or community environment, or for other reasons such as illness or emotional disturbance.

The extent to which such factors can affect an individual's psychological development provides an important problem for investigation. This question, however, should not be confused with that of the reliability of a particular test. When we measure the reliability of the Stanford-Binet, for example, we do not ordinarily correlate retest scores over a period of ten years, or even one year, but over a few weeks. To be sure, long-range retests have been conducted with such tests, but the results are generally discussed in terms of the predictability of adult intelligence from childhood performance, rather than in terms of the reliability of a particular test. The concept of reliability is generally restricted to short-range, random changes that characterize the test performance itself rather than the entire behavior domain that is being tested.

It should be noted that different behavior functions may themselves vary in the extent of daily fluctuation they exhibit. For example, steadiness of delicate finger movements is undoubtedly more susceptible to slight changes in the person's condition than is verbal comprehension. If we wish to obtain an over-all estimate of the individual's habitual finger steadiness, we would probably require repeated tests on several days, whereas a single test session would suffice for verbal comprehension. Again we must fall back on an analysis of the purposes of the test and on a thorough understanding of the behavior the test is designed to predict.

Although apparently simple and straightforward, the test-retest technique presents difficulties when applied to most psychological tests. Practice will probably produce varying amounts of improvement in the retest scores of different individuals. Moreover, if the interval between retests is fairly short, the examinees may recall many of their former responses. In other words, the same pattern of right and wrong responses is likely to recur through sheer memory. Thus, the scores on the two administrations of the test are not independently obtained and the correlation between them will be spuriously high. The nature of the test itself may also change with repetition. This is especially true of problems involving reasoning or ingenuity. Once the subject has grasped the principle involved in the problem, or once he has worked out a solution, he can reproduce the correct response in the future without going through the intervening steps. Only tests that are not appreciably affected by repetition lend themselves to the retest technique. A number of sensory discrimination and motor tests would fall into this category. For the large majority of psychological tests, however, the retest technique is inappropriate.

ALTERNATE-FORM RELIABILITY. One way of avoiding the difficulties encountered in test-retest reliability is through the use of alternate forms of the test. The same persons can thus be tested with one form on the first occasion and with another, comparable form on the second. The correlation between the scores obtained on the two forms represents the reliability coefficient of the test. It will be noted that such a reliability coefficient is a measure of both temporal stability and consistency of response to different item samples (or test forms). This coefficient thus combines two types of reliability. Since both types are important for most testing purposes, however, alternate-form reliability provides a useful measure for evaluating many tests.

The concept of item sampling, or *content sampling,* underlies not only alternate-form reliability but also other types of reliability to be discussed shortly. It is therefore appropriate to examine it more closely. Everyone has probably had the experience of taking a course examination in which he felt he had a "lucky break" because many of the items covered the very topics he happened to have studied most carefully. On another occasion, he may have had the opposite experience, finding an unusually large number of items on areas he had failed to review. This familiar situation illustrates error variance resulting from content sampling. To what extent do scores on this test depend on factors *specific* to the particular selection of items? If a different investigator, working independently, were to prepare another test in accordance with the same specifications, how much would an individual's score differ on the two tests?

Let us suppose that a 40-item vocabulary test has been constructed as a measure of general verbal comprehension. Now suppose that a second list of 40 different words is assembled for the same purpose, and that the items are constructed with equal care to cover the same range of difficulty as the first test. The differences in the scores obtained by the same individuals on these two tests illustrate the type of error variance under consideration. Owing to fortuitous factors in the past experience of different individuals, the relative difficulty of the two lists will vary somewhat from person to person. Thus, the first list might contain a larger number of words unfamiliar to individual A than does the second list. The second list, on the other hand, might contain a disproportionately large number of words unfamiliar to individual B. If the two individuals are approximately equal in their overall word knowledge (i.e., in their "true scores"), B will nevertheless excel A on the first list, while A will excel B on the second. The relative standing of these two persons will therefore be reversed on the two lists, owing to chance differences in the selection of items.

Like test-retest reliability, alternate-form reliability should always be accompanied by a statement of the length of the interval between test administrations, as well as a description of relevant intervening experiences. If the two forms are administered *in immediate succession,* the resulting correlation shows reliability across forms only, not across occasions. The error variance in this case represents fluctuations in performance from one set of items to another, but not fluctuations over time.

In the development of alternate forms, care should of course be exercised to ensure that they are truly parallel. Fundamentally, parallel forms of a test should be independently constructed tests designed to meet the same specifications. The tests should contain the same number of items, and the items should be expressed in the same form and should cover the same type of content. The range and level of difficulty of the items should also be equal. Instructions, time limits, illustrative examples, format, and all other aspects of the test must likewise be checked for comparability.

It should be added that the availability of parallel test forms is desirable for other reasons besides the determination of test reliability. Alternate forms are useful in follow-up studies or in investigations of the effects of some intervening experimental factor on test performance. The use of several alternate forms also provides a means of reducing the possibility of coaching or cheating.

Although much more widely applicable than test-retest reliability, alternate-form reliability also has certain limitations. In the first place, if the behavior functions under consideration are subject to a large practice effect, the use of alternate forms will reduce but not eliminate such an effect. To be sure, if all examinees were to show the same improvement with repetition, the correlation between their scores would remain unaffected, since adding a constant amount to each score does not alter the correlation coefficient. It is much more likely, however, that individuals will differ in amount of improvement, owing to extent of previous practice with similar material, motivation in taking the test, and other factors. Under these conditions, the practice effect represents another source of variance that will tend to reduce the correlation between the two test forms. If the practice effect is small, reduction will be negligible.

Another related question concerns the degree to which the nature of the test will change with repetition. In certain types of ingenuity problems, for example, any item involving the same principle can be readily solved by most subjects once they have worked out the solution to the first. In such a case, changing the specific content of the items in the second form would not suffice to eliminate this carry-over from the first form. Finally, it should be added that alternate forms are unavailable for many tests, because of the practical difficulties of constructing comparable forms. For all these reasons, other techniques for estimating test reliability are often required.

SPLIT-HALF RELIABILITY. From a single administration of one form of a test it is possible to arrive at a measure of reliability by various split-half procedures. In such a way, two scores are obtained for each person by dividing the test into comparable halves. It is apparent that split-half reliability provides a measure of consistency with regard to content sampling. Temporal stability of the scores does not enter into such reliability, because only one test session is involved. This type of reliability coefficient is sometimes called a coefficient of internal consistency, since only a single administration of a single form is required.

To find split-half reliability, the first problem is how to split the test in order to obtain the most nearly comparable halves. Any test can be divided in many different ways. In most tests, the first half and the second half would not be comparable, owing to differences in nature and difficulty level of items, as well as to the cumulative effects of warming up, practice, fatigue, boredom, and any other factors varying progressively from the beginning to the end of the test. A procedure that is adequate for most purposes is to find the scores on the odd and even items of the test. If the items were originally arranged in an approximate order of difficulty, such a division yields very nearly equivalent half-scores. One precaution to be observed in making such an odd-even split pertains to groups of items dealing with a single problem, such as questions referring to a particular mechanical diagram or to a given passage in a reading test. In this case, a whole group of items should be assigned intact to one or the other half. Were the items in such a group to be placed in different halves of the test, the similarity of the half-scores would be spuriously inflated, since any single error in understanding of the problem might affect items in both halves.

Once the two half-scores have been obtained for each person, they may be correlated by the usual method. It should be noted, however, that this correlation actually gives the reliability of only a half-test. For example, if the entire test consists of 100 items, the correlation is computed between two sets of scores each of which is based on only 50 items. In both test-retest and alternate-form reliability, on the other hand, each score is based on the full number of items in the test.

Other things being equal, the longer a test, the more reliable it will be.[2] It is reasonable to expect that, with a larger sample of behavior, we can arrive at a more adequate and consistent measure. The effect that lengthening or shortening a test will have on its coefficient can be estimated by means of the Spearman-Brown formula, given below:

$$r_{11} = \frac{nr'_{11}}{1 + (n - 1)r'_{11}}$$

in which r_{11} is the estimated coefficient, r'_{11} the obtained coefficient, and n is the number of times the test is lengthened or shortened. Thus, if the number of test items is increased from 25 to 100, n is 4; if it is decreased from 60 to 30, n is $\frac{1}{2}$. The Spearman-Brown formula is widely used in determining reliability by the split-half method, many test manuals reporting reliability in this form. When applied to split-half reliability, the formula always involves doubling the length of the test. Under these conditions, it can be simplified as follows:

---

[2] Lengthening a test, however, will increase only its consistency in terms of content sampling, not its stability over time (see Cureton, 1965).

$$r_{11} = \frac{2r'_{11}}{1 + r'_{11}}$$

An alternate method for finding spit-half reliability was developed by Rulon (1939). It requires only the variance of the *differences* between each person's scores on the two half-tests (σ^2_d) and the variance of total scores (σ^2_x); these two values are substituted in the following formula, which yields the reliability of the whole test directly:

$$r_{11} = 1 - \frac{\sigma^2_d}{\sigma^2_x}$$

It is interesting to note the relationship of this formula to the definition of error variance. Any difference between a person's scores on the two half-tests represents chance error. The variance of these differences, divided by the variance of total scores, gives the proportion of error variance in the scores. When this error variance is subtracted from 1.00, it gives the proportion of "true" variance, which is equal to the reliability coefficient.

KUDER-RICHARDSON RELIABILITY. A fourth method for finding reliability, also utilizing a single administration of a single form, is based on the consistency of responses to all items in the test. This *interitem consistency* is influenced by two sources of error variance: (1) content sampling (as in alternate-form and split-half reliability); and (2) heterogeneity of the behavior domain sampled. The more homogeneous the domain, the higher the interitem consistency. For example, if one test includes only multiplication items, while another comprises addition, subtraction, multiplication, and division items, the former test will probably show more interitem consistency than the latter. In the latter, more heterogeneous test, one subject may perform better in subtraction than in any of the other arithmetic operations; another subject may score relatively well on the division items, but more poorly in addition, subtraction, and multiplication; and so on. A more extreme example would be represented by a test consisting of 40 vocabulary items, in contrast to one containing 10 vocabulary, 10 spatial relations, 10 arithmetic reasoning, and 10 perceptual speed items. In the latter test, there might be little or no relationship between an individual's performance on the different types of items.

It is apparent that test scores will be less ambiguous when derived from relatively homogeneous tests. Suppose that in the highly heterogeneous, 40-item test cited above, Smith and Jones both obtain a score of 20. Can we conclude that the performances of the two on this test were equal? Not at all. Smith may have correctly completed 10 vocabulary items, 10 perceptual speed items, and none of the arithmetic reasoning and spatial relations items. In contrast, Jones may have received a score of 20 by the successful completion of 5 perceptual speed, 5 spatial relations, 10 arithmetic reasoning, and no vocabulary items.

Many other combinations could obviously produce the same total score of 20. This score would have a very different meaning when obtained through such dissimilar combinations of items. In the relatively homogeneous vocabulary test, on the other hand, a score of 20 would probably mean that the subject had succeeded with approximately the first 20 words, if the items were arranged in ascending order of difficulty. He might have failed two or three easier words and correctly responded to two or three more difficult items beyond the 20th, but such individual variations are slight in comparison with those found in a more heterogeneous test.

A highly relevant question in this connection is whether the criterion that the test is trying to predict is itself relatively homogeneous or heterogeneous. Although homogeneous tests are to be preferred because their scores permit fairly unambiguous interpretation, a single homogeneous test is obviously not an adequate predictor of a highly heterogeneous criterion. Moreover, in the prediction of a heterogeneous criterion, the heterogeneity of test items would not necessarily represent error variance. Traditional intelligence tests provide a good example of heterogeneous tests designed to predict heterogeneous criteria. In such a case, however, it may be desirable to construct several relatively homogeneous tests, each measuring a different phase of the heterogeneous criterion. Thus, unambiguous interpretation of test scores could be combined with adequate criterion coverage.

The most common procedure for finding interitem consistency is that developed by Kuder and Richardson (1937). As in the split-half methods, interitem consistency is found from a single administration of a single test. Rather than requiring two half-scores, however, such a technique is based on an examination of performance on each item. Of the various formulas derived in the original article, the most widely applicable, commonly known as "Kuder-Richardson formula 20," is the following:[3]

$$r_{11} = \left(\frac{n}{n-1}\right)\frac{\sigma^2_t - \Sigma pq}{\sigma^2_t}$$

In this formula, r_{11} is the reliability coefficient of the whole test, n is the number of items in the test, and σ_t the standard deviation of total scores on the test. The only new term in this formula, Σpq, is found by tabulating the proportion of persons who pass (p) and the proportion who do not pass (q) each item. The product of p and q is computed for each item, and these products are then added for all items, to give Σpq. Since in the process of test construction p is often routinely recorded in order to find the difficulty level of each item, this method of determining reliability involves little additional computation.

---

[3] A simple derivation of this formula can be found in Ebel (1965, pp. 320–327).

It can be shown mathematically that the Kuder-Richardson reliability coefficient is actually the mean of all split-half coefficients resulting from different splittings of a test (Cronbach, 1951).[4] The ordinary split-half coefficient, on the other hand, is based on a planned split designed to yield equivalent sets of items. Hence, unless the test items are highly homogeneous, the Kuder-Richardson coefficient will be lower than the split-half reliability. An extreme example will serve to highlight the difference. Suppose we construct a 50-item test out of 25 different kinds of items such that items 1 and 2 are vocabulary items, items 3 and 4 arithmetic reasoning, items 5 and 6 spatial orientation, and so on. The odd and even scores on this test could theoretically agree quite closely, thus yielding a high split-half reliability coefficient. The homogeneity of this test, however, would be very low, since there would be little consistency of performance among the entire set of 50 items. In this example, we would expect the Kuder-Richardson reliability to be much lower than the split-half reliability. It can be seen that the difference between Kuder-Richardson and split-half reliability coefficients may serve as a rough index of the heterogeneity of a test.

The Kuder-Richardson formula is applicable to tests whose items are scored as right or wrong, or according to some other all-or-none system. Some tests, however, may have multiple-scored items. On a personality inventory, for example, the respondent may receive a different numerical score on an item, depending on whether he checks "usually," "sometimes," "rarely," or "never." For such tests, a generalized formula has been derived, known as coefficient alpha (Cronbach, 1951; Novick & Lewis, 1967). In this formula, the value Σpq is replaced by $\Sigma\sigma^2_i$, the sum of the variances of item scores. The procedure is to find the variance of all individuals' scores for each item and then to add these variances across all items. The complete formula for coefficient alpha is given below:

$$r_{11} = \left(\frac{n}{n-1}\right)\frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_t}$$

A clear description of the computational layout for finding coefficient alpha can be found in Ebel (1965, pp. 326–330).

SCORER RELIABILITY. It should now be apparent that the different types of reliability vary in the factors they subsume under error variance. In one case, error variance covers temporal fluctuations; in another, it refers to differences between sets of parallel items; and in still another, it includes any interitem inconsistency. On the other hand, the factors *excluded* from measures of error variance are broadly of two types: (*a*) those factors whose variance should remain in the scores, since they are part of the true differences under consideration; and (*b*) those irrelevant factors that can be experimentally controlled. For example, it is not customary to report the error of measurement resulting when a test is administered under distracting conditions or with a longer or shorter time limit than that specified in the manual. Timing errors and serious distractions can be empirically eliminated from the testing situation. Hence, it is not necessary to report special reliability coefficients corresponding to "distraction variance" or "timing variance."

---

[4] This is strictly true only when the split-half coefficients are found by the Rulon formula, not when they are found by correlation of halves and Spearman-Brown formula (Novick & Lewis, 1967).

Similarly, most tests provide such highly standardized procedures for administration and scoring that error variance attributable to these factors is negligible. This is particularly true of group tests designed for mass testing and computer scoring. With such instruments, we need only to make certain that the prescribed procedures are carefully followed and adequately checked. With clinical instruments employed in intensive individual examinations, on the other hand, there is evidence of considerable "examiner variance." Through special experimental designs, it is possible to separate this variance from that attributable to temporal fluctuations in the subject's condition or to the use of alternate test forms.

One source of error variance that can be checked quite simply is scorer variance. Certain types of tests—notably tests of creativity and projective tests of personality—leave a good deal to the judgment of the scorer. With such tests, there is as much need for a measure of scorer reliability as there is for the more usual reliability coefficients. Scorer reliability can be found by having a sample of test papers independently scored by two examiners. The two scores thus obtained by each examinee are then correlated in the usual way, and the resulting correlation coefficient is a measure of scorer reliability. This type of reliability is commonly computed when subjectively scored instruments are employed in research. Test manuals should also report it when appropriate.

OVERVIEW. The different types of reliability coefficients discussed in this section are summarized in Tables 8 and 9. In Table 8 the operations followed in obtaining each type of reliability are classified with regard to number of test forms and number of testing sessions required. Table 9 shows the sources of variance treated as error variance by each procedure.

Any reliability coefficient may be interpreted directly in terms of the *percentage of score variance* attributable to different sources. Thus, a reliability coefficient of .85 signifies that 85 percent of the variance in test

Table 8
Techniques for Measuring Reliability, in Relation to Test Form and Testing Session

Testing Sessions Required	Test Forms Required	
	One	Two
One	Split-Half Kuder-Richardson Scorer	Alternate-Form (Immediate)
Two	Test-Retest	Alternate-Form (Delayed)

scores depends on true variance in the trait measured and 15 percent depends on error variance (as operationally defined by the specific procedure followed). The statistically sophisticated reader may recall that it is the *square* of a correlation coefficient that represents proportion of common variance. Actually, the proportion of true variance in test scores is the square of the correlation between scores on a single form of the test and true scores, free from chance errors. This correlation, known as the index of reliability,[5] is equal to the square root of the reliability co-

Table 9
Sources of Error Variance in Relation to Reliability Coefficients

Type of Reliability Coefficient	Error Variance
Test-Retest	Time sampling
Alternate-Form (Immediate)	Content sampling
Alternate-Form (Delayed)	Time sampling and Content sampling
Split-Half	Content sampling
Kuder-Richardson and Coefficient Alpha	Content sampling and Content heterogeneity
Scorer	Interscorer differences

---

[5] Derivations of the index of reliability, based on two different sets of assumptions, are given by Gulliksen (1950b, Chs. 2 and 3).

efficient ($\sqrt{r_{11}}$). When the index of reliability is squared, the result is the reliability coefficient (r_{11}), which can therefore be interpreted directly as the percentage of true variance.

Experimental designs that yield more than one type of reliability coefficient for the same group permit the analysis of total score variance into different components. Let us consider the following hypothetical example. Forms A and B of a creativity test have been administered with a two-month interval to 100 sixth-grade children. The resulting alternate-form reliability is .70. From the responses of either form, a split-half reliability coefficient can also be computed.[6] This coefficient, stepped up by the Spearman-Brown formula, is .80. Finally, a second scorer has rescored a random sample of 50 papers, from which a scorer reliability of .92 is obtained. The three reliability coefficients can now be analyzed to yield the error variances shown in Table 10 and Figure 11. It will be noted that by subtracting the error variance attributable to content sampling alone (split-half reliability) from the error variance attributable to both content and time sampling (alternate-form reliability), we find that .10 of the variance can be attributed to time sampling alone. Adding the error variances attributable to content sampling (.20), time sampling (.10), and interscorer difference (.08) gives a total error variance of .38 and hence a true variance of .62. These proportions, expressed in the more familiar percentage terms, are shown graphically in Figure 11.

TABLE 10

Analysis of Sources of Error Variance in a Hypothetical Test

From delayed alternate-form reliability:	$1 - .70 = .30$	(time sampling plus content sampling)
From split-half, Spearman-Brown reliability:	$1 - .80 = .20^*$	(content sampling)
Difference	$.10^*$	(time sampling)
From scorer reliability:	$1 - .92 = .08^*$	(interscorer difference)
Total Measured Error Variance*	$= .20 + .10 + .08 = .38$	
True Variance	$= 1 - .38 = .62$	

---

[6] For a better estimate of the coefficient of internal consistency, split-half correlations could be computed for each form and the two coefficients averaged by the appropriate statistical procedures.

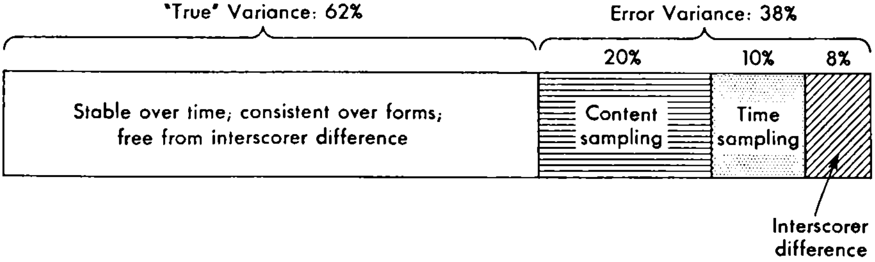


FIG. 11. Percentage Distribution of Score Variance in a Hypothetical Test.

RELIABILITY OF SPEEDED TESTS

Both in test construction and in the interpretation of test scores, an important distinction is that between the measurement of speed and of power. A pure *speed test* is one in which individual differences depend entirely on speed of performance. Such a test is constructed from items of uniformly low difficulty, all of which are well within the ability level of the persons for whom the test is designed. The time limit is made so short that no one can finish all the items. Under these conditions, each person's score reflects only the speed with which he worked. A pure *power test,* on the other hand, has a time limit long enough to permit everyone to attempt all items. The difficulty of the items is steeply graded, and the test includes some items too difficult for anyone to solve, so that no one can get a perfect score.

It will be noted that both speed and power tests are designed to prevent the achievement of perfect scores. The reason for such a precaution is that perfect scores are indeterminate, since it is impossible to know how much higher the individual's score would have been if more items, or more difficult items, had been included. To enable each individual to show fully what he is able to accomplish, the test must provide adequate ceiling, either in number of items or in difficulty level. An exception to this rule is found in mastery testing, as illustrated by the criterion-referenced tests discussed in Chapter 4. The purpose of such testing is not to establish the limits of what the individual can do, but to determine whether a preestablished performance level has or has not been reached.

In actual practice, the distinction between speed and power tests is one of degree, most tests depending on both power and speed in varying proportions. Information about these proportions is needed for each test in order not only to understand what the test measures but also to choose the proper procedures for evaluating its reliability. Single-trial reliability coefficients, such as those found by odd-even or Kuder-Richardson techniques, are inapplicable to speeded tests. To the extent that individual differences in test scores depend on speed of performance, reliability coefficients found by these methods will be spuriously high. An extreme example will help to clarify this point. Let us suppose that a 50-item test depends entirely on speed, so that individual differences in score are based wholly on number of items attempted, rather than on errors. Then, if individual A obtains a score of 44, he will obviously have 22 correct odd items and 22 correct even items. Similarly, individual B, with a score of 34, will have odd and even scores of 17 and 17, respectively. Consequently, except for accidental careless errors on a few items, the correlation between odd and even scores would be perfect, or +1.00. Such a correlation, however, is entirely spurious and provides no information about the reliability of the test.

An examination of the procedures followed in finding both split-half and Kuder-Richardson reliability will show that both are based on the consistency in *number of errors* made by the examinee. If, now, individual differences in test scores depend, not on errors, but on speed, the measure of reliability must obviously be based on consistency in *speed of work.* When test performance depends on a combination of speed and power, the single-trial reliability coefficient will fall below 1.00, but it will still be spuriously high. As long as individual differences in test scores are appreciably affected by speed, single-trial reliability coefficients cannot be properly interpreted.

What alternative procedures are available to determine the reliability of significantly speeded tests? If the test-retest technique is applicable, it would be appropriate. Similarly, equivalent-form reliability may be properly employed with speed tests. Split-half techniques may also be used, provided that the split is made in terms of time rather than in terms of items. In other words, the half-scores must be based on separately timed parts of the test. One way of effecting such a split is to administer two equivalent halves of the test with separate time limits. For example, the odd and even items may be separately printed on different pages, and each set of items given with one-half the time limit of the entire test. Such a procedure is tantamount to administering two equivalent forms of the test in immediate succession. Each form, however, is half as long as the test proper, while the subjects' scores are normally based on the whole test. For this reason, either the Spearman-Brown or some other appropriate formula should be used to find the reliability of the whole test.

If it is not feasible to administer the two half-tests separately, an alternative procedure is to divide the total time into quarters, and to find a score for each of the four quarters. This can easily be done by having the examinees mark the item on which they are working whenever the examiner gives a prearranged signal. The number of items correctly completed within the first and fourth quarters can then be combined to represent one half-score, while those in the second and third quarters can be combined to yield the other half-score. Such a combination of quarters tends to balance out the cumulative effects of practice, fatigue, and other factors. This method is especially satisfactory when the items are not steeply graded in difficulty level.

When is a test appreciably speeded? Under what conditions must the special precautions discussed in this section be observed? Obviously, the mere employment of a time limit does not signify a speed test. If all subjects finish within the given time limit, speed of work plays no part in determining the scores. Percentage of persons who fail to complete the test might be taken as a crude index of speed versus power. Even when no one finishes the test, however, the role of speed may be negligible. For example, if everyone completes exactly 40 items of a 50-item test, individual differences with regard to speed are entirely absent, although no one had time to attempt all the items.

The essential question, of course, is: "To what extent are individual differences in test scores attributable to speed?" In more technical terms, we want to know what proportion of the total variance of test scores is speed variance. This proportion can be estimated roughly by finding the variance of number of items completed by different persons and dividing it by the variance of total test scores (σ^2_c/σ^2_t). In the example cited above, in which every individual finishes 40 items, the numerator of this fraction would be zero, since there are no individual differences in number of items completed ($\sigma^2_c = 0$). The entire index would thus equal zero in a pure power test. On the other hand, if the total test variance (σ^2_t) is attributable to individual differences in speed, the two variances will be equal and the ratio will be 1.00. Several more refined procedures have been developed for determining this proportion, but their detailed consideration falls beyond the scope of this book.[7]

An example of the effect of speed on single-trial reliability coefficients is provided by data collected in an investigation of the first edition of the SRA Tests of Primary Mental Abilities for Ages 11 to 17 (Anastasi & Drake, 1954). In this study, the reliability of each test was first determined by the usual odd-even procedure. These coefficients, given in the first row of Table 11, are closely similar to those reported in the test manual. Reliability coefficients were then computed by correlating scores on separately timed halves. These coefficients are shown in the second row of Table 11. Calculation of speed indexes showed that the Verbal Meaning test is primarily a power test, while the Reasoning test is somewhat more dependent on speed. The Space and Number tests proved to be highly speeded. It will be noted in Table 11 that, when properly com-

---

[7] See, e.g., Cronbach & Warrington (1951), Gulliksen (1950a, 1950b), Guttman (1955), Helmstadter & Ortmeyer (1953).

Table 11
Reliability Coefficients of Four of the SRA Tests of Primary Mental Abilities for Ages 11 to 17 (1st Edition)

(Data from Anastasi & Drake, 1954)

Reliability Coefficient Found by:	Verbal Meaning	Reasoning	Space	Number
Single-trial odd-even method	.94	.96	.90	.92
Separately timed halves	.90	.87	.75	.83

puted, the reliability of the Space test is .75, in contrast to a spuriously high odd-even coefficient of .90. Similarly, the reliability of the Reasoning test drops from .96 to .87, and that of the Number test drops from .92 to .83. The reliability of the relatively unspeeded Verbal Meaning test, on the other hand, shows a negligible difference when computed by the two methods.

DEPENDENCE OF RELIABILITY COEFFICIENTS ON THE SAMPLE TESTED

HETEROGENEITY. An important factor influencing the size of a reliability coefficient is the nature of the group on which reliability is measured. In the first place, any correlation coefficient is affected by the range of individual differences in the group. If every member of a group were alike in spelling ability, then the correlation of spelling with any other ability would be zero in that group. It would obviously be impossible, within such a group, to predict an individual's standing in any other ability from a knowledge of his spelling score.

Another, less extreme, example is provided by the correlation between two aptitude tests, such as a verbal comprehension and an arithmetic reasoning test. If these tests were administered to a highly homogeneous sample, such as a group of 300 college sophomores, the correlation between the two would probably be close to zero. There is little relationship, within such a selected sample of college students, between any individual's verbal ability and his numerical reasoning ability. On the other hand, were the tests to be given to a heterogeneous sample of 300 persons, ranging from institutionalized mentally retarded persons to college graduates, a high correlation would undoubtedly be obtained between the two tests. The mentally retarded would obtain poorer scores than the college graduates on *both* tests, and similar relationships would hold for other subgroups within this highly heterogeneous sample.

Examination of the hypothetical scatter diagram given in Figure 12 will further illustrate the dependence of correlation coefficients on the variability, or extent of individual differences, within the group. This scatter diagram shows a high positive correlation in the entire, heterogeneous group, since the entries are closely clustered about the diagonal extending from lower left- to upper right-hand corners. If, now, we consider only the subgroup falling within the small rectangle in the upper right-hand portion of the diagram, it is apparent that the correlation between the two variables is close to zero. Individuals falling within this restricted range in both variables represent a highly homogeneous group, as did the college sophomores mentioned above.

Like all correlation coefficients, reliability coefficients depend on the variability of the sample within which they are found. Thus, if the reliability coefficient reported in a test manual was determined in a group ranging from fourth-grade children to high school students, it cannot be assumed that the reliability would be equally high within, let us say, an eighth-grade sample. When a test is to be used to discriminate individual

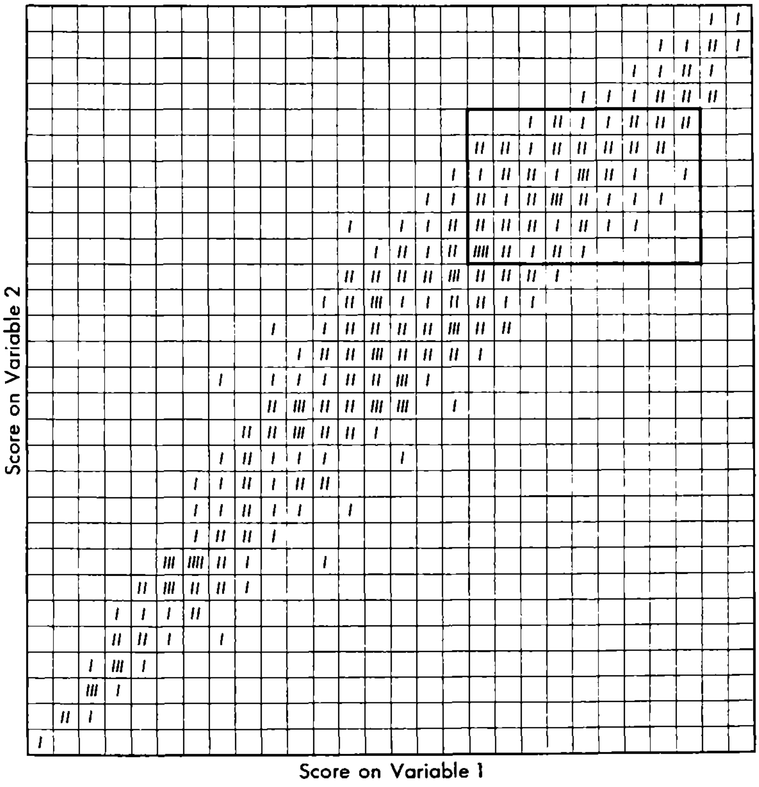


FIG. 12. The Effect of Restricted Range upon a Correlation Coefficient.

differences within a more homogeneous sample than the standardization group, the reliability coefficient should be redetermined on such a sample. Formulas for estimating the reliability coefficient to be expected when the standard deviation of the group is increased or decreased are available in elementary statistics textbooks. It is preferable, however, to recompute the reliability coefficient empirically on a group comparable to that on which the test is to be used. For tests designed to cover a wide range of age or ability, the test manual should report separate reliability coefficients for relatively homogeneous subgroups within the standardization sample.

ABILITY LEVEL. Not only does the reliability coefficient vary with the extent of individual differences in the sample, but it may also vary between groups differing in average ability level. These differences, moreover, cannot usually be predicted or estimated by any statistical formula, but can be discovered only by empirical tryout of the test on groups differing in age or ability level. Such differences in the reliability of a single test may arise from the fact that a slightly different combination of abilities is measured at different difficulty levels of the test. Or it may result from the statistical properties of the scale itself, as in the Stanford-Binet (Pinneau, 1961, Ch. 5). Thus, for different ages and for different IQ levels, the reliability coefficient of the Stanford-Binet varies from .83 to .98. In other tests, reliability may be relatively low for the younger and less able groups, since their scores are unduly influenced by guessing. Under such circumstances, the particular test should not be employed at these levels.

It is apparent that every reliability coefficient should be accompanied by a full description of the type of group on which it was determined. Special attention should be given to the variability and the ability level of the sample. The reported reliability coefficient is applicable only to samples similar to that on which it was computed. A desirable and growing practice in test construction is to fractionate the standardization sample into more homogeneous subgroups, with regard to age, sex, grade level, occupation, and the like, and to report separate reliability coefficients for each subgroup. Under these conditions, the reliability coefficients are more likely to be applicable to the samples with which the test is to be used in actual practice.

STANDARD ERROR OF MEASUREMENT

INTERPRETATION OF INDIVIDUAL SCORES. The reliability of a test may be expressed in terms of the standard error of measurement ($\sigma_{meas.}$), also called the standard error of a score. This measure is particularly well suited to the interpretation of individual scores. For many testing purposes, it is therefore more useful than the reliability coefficient. The standard error of measurement can be easily computed from the reliability coefficient of the test, by the following formula:

$$\sigma_{meas.} = \sigma_1\sqrt{1 - r_{11}}$$

in which σ_1 is the standard deviation of the test scores and r_{11} the reliability coefficient, both computed on the same group. For example, if deviation IQ's on a particular intelligence test have a standard deviation of 15 and a reliability coefficient of .89, the $\sigma_{meas.}$ of an IQ on this test is: $15\sqrt{1 - .89} = 15\sqrt{.11} = 15(.33) = 5$.

To understand what the $\sigma_{meas.}$ tells us about a score, let us suppose that we had a set of 100 IQ's obtained with the above test by a single boy, Jim. Because of the types of chance errors discussed in this chapter, these scores will vary, falling into a normal distribution around Jim's true score. The mean of this distribution of 100 scores can be taken as the true score and the standard deviation of the distribution can be taken as the $\sigma_{meas.}$. Like any standard deviation, this standard error can be interpreted in terms of the normal curve frequencies discussed in Chapter 4 (see Figure 3). It will be recalled that between the mean and $\pm 1\sigma$ there are approximately 68 percent of the cases in a normal curve. Thus, we can conclude that the chances are roughly 2:1 (or 68:32) that Jim's IQ on this test will fluctuate between $\pm 1\sigma_{meas.}$ or 5 points on either side of his true IQ. If his true IQ is 110, we would expect him to score between 105 and 115 about two-thirds (68 percent) of the time.

If we want to be more certain of our prediction, we can choose higher odds than 2:1. Reference to Figure 3 in Chapter 4 shows that $\pm 3\sigma$ covers 99.7 percent of the cases. It can be ascertained from normal curve frequency tables that a distance of 2.58σ on either side of the mean includes exactly 99 percent of the cases. Hence, the chances are 99:1 that Jim's IQ will fall within $2.58\sigma_{meas.}$, or $(2.58)(5) = 13$ points, on either side of his true IQ. We can thus state at the 99 percent confidence level (with only one chance of error out of 100) that Jim's IQ on any single administration of the test will lie between 97 and 123 ($110 - 13$ and $110 + 13$). If Jim were given 100 equivalent tests, his IQ would fall outside this band of values only once.

In actual practice, of course, we do not have the true scores, but only the scores obtained in a single test administration. Under these circumstances, we could try to follow the above reasoning in the reverse direction. If an individual's obtained score is unlikely to deviate by more than $2.58\sigma_{meas.}$ from his true score, we could argue that his *true* score must lie within $2.58\sigma_{meas.}$ of his *obtained* score. Although we cannot assign a probability to this statement for any given obtained score, we *can* say that the statement would be correct for 99 percent of all the cases. On the basis of this reasoning, Gulliksen (1950b, pp. 17–20) proposed that the standard error of measurement be used as illustrated above to estimate the reasonable limits of the true score for persons with any given obtained score. It is in terms of such "reasonable limits" that the error of measurement is customarily interpreted in psychological testing and it will be so interpreted in this book.

The standard error of measurement and the reliability coefficient are obviously alternative ways of expressing test reliability. Unlike the reliability coefficient, the error of measurement is independent of the variability of the group on which it is computed. Expressed in terms of individual scores, it remains unchanged when found in a homogeneous or a heterogeneous group. On the other hand, being reported in score units, the error of measurement will not be directly comparable from test to test. The usual problems of comparability of units would thus arise when errors of measurement are reported in terms of arithmetic problems, words in a vocabulary test, and the like. Hence, if we want to compare the reliability of *different tests,* the reliability coefficient is the better measure. To interpret *individual scores,* the standard error of measurement is more appropriate.

INTERPRETATION OF SCORE DIFFERENCES. It is particularly important to consider test reliability and errors of measurement when evaluating the *differences* between two scores. Thinking in terms of the range within which each score may fluctuate serves as a check against overemphasizing small differences between scores. Such caution is desirable both when comparing test scores of different persons and when comparing the scores of the same individual in different abilities. Similarly, changes in scores following instruction or other experimental variables need to be interpreted in the light of errors of measurement.

A frequent question about test scores concerns the individual's relative standing in different areas. Is Jane more able along verbal than along numerical lines? Does Tom have more aptitude for mechanical than for verbal activities? If Jane scored higher on the verbal than on the numerical subtests on an aptitude battery and Tom scored higher on the mechanical than on the verbal, how sure can we be that they would still do so on a retest with another form of the battery? In other words, could the score differences have resulted merely from the chance selection of specific items in the particular verbal, numerical, and mechanical tests employed?

Because of the growing interest in the interpretation of score profiles, test publishers have been developing report forms that permit the evalua-

	Verbal Reasoning	Numerical Ability	VR + NA	Abstract Reasoning	Clerical Sp. & Acc.	Mechanical Reasoning	Space Relations	Spelling	Language Usage
RAW SCORE	21	30	51	43	38	44	39	96	40
PERCENTILE	60	95	80	95	30	80	90	99	85

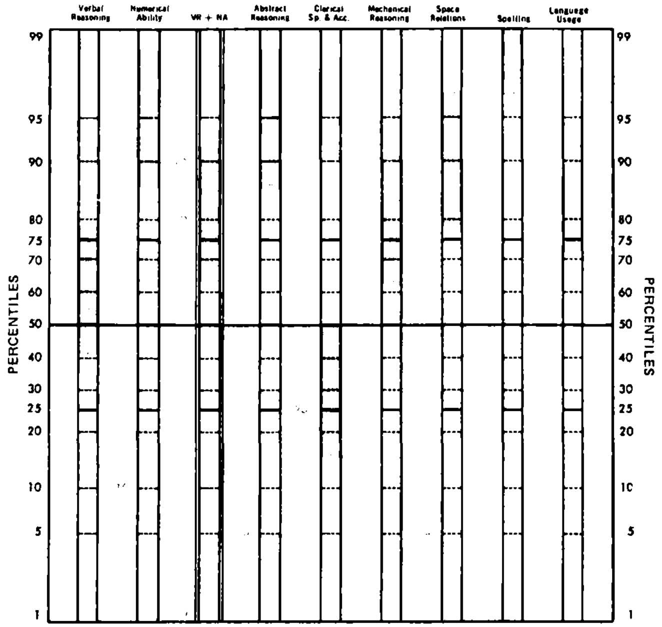


FIG. 13. Score Profile on the Differential Aptitude Tests, Illustrating Use of Percentile Bands.

(Fig. 2, Fifth Edition Manual, p. 73. Reproduced by permission. Copyright © 1973, 1974 by The Psychological Corporation, New York, N.Y. All rights reserved.)

tion of scores in terms of their errors of measurement. An example is the Individual Report Form for use with the Differential Aptitude Tests, reproduced in Figure 13. On this form, percentile scores on each subtest of the battery are plotted as one-inch bars, with the obtained percentile at the center. Each percentile bar corresponds to a distance of approximately 1½ to 2 standard errors on either side of the obtained score.[8] Hence the assumption that the individual's "true" score falls within the bar is correct about 90 percent of the time. In interpreting the profiles, test users are advised not to attach importance to differences between scores whose percentile bars overlap, especially if they overlap by more than half their length. In the profile illustrated in Figure 13, for example, the difference between the Verbal Reasoning and Numerical Ability scores probably reflects a genuine difference in ability level; that between Mechanical Reasoning and Space Relations probably does not; the difference between Abstract Reasoning and Mechanical Reasoning is in the doubtful range.

---

[8] Because the reliability coefficient (and hence the $\sigma_{meas.}$) varies somewhat with subtest, grade, and sex, the actual ranges covered by the one-inch lines are not identical, but they are sufficiently close to permit uniform interpretations for practical purposes.

It is well to bear in mind that the standard error of the difference between two scores is larger than the error of measurement of either of the two scores. This follows from the fact that this difference is affected by the chance errors present in *both* scores. The standard error of the difference between two scores can be found from the standard errors of measurement of the two scores by the following formula:

$$\sigma_{diff.} = \sqrt{\sigma^2_{meas._1} + \sigma^2_{meas._2}}$$

in which $\sigma_{diff.}$ is the standard error of the difference between the two scores, and $\sigma_{meas._1}$ and $\sigma_{meas._2}$ are the standard errors of measurement of the separate scores. By substituting $SD\sqrt{1 - r_{1I}}$ for $\sigma_{meas._1}$ and $SD\sqrt{1 - r_{2II}}$ for $\sigma_{meas._2}$, we may rewrite the formula directly in terms of reliability coefficients, as follows:

$$\sigma_{diff.} = SD\sqrt{2 - r_{1I} - r_{2II}}$$

In this substitution, the same *SD* was used for tests 1 and 2, since their scores would have to be expressed in terms of the same scale before they could be compared.

We may illustrate the above procedure with the Verbal and Performance IQ's on the Wechsler Adult Intelligence Scale (WAIS). The split-half reliabilities of these scores are .96 and .93, respectively. WAIS deviation IQ's have a mean of 100 and an *SD* of 15. Hence the standard error of the difference between these two scores can be found as follows:

$$\sigma_{diff.} = 15\sqrt{2 - .96 - .93} = 4.95$$

To determine how large a score difference could be obtained by chance at the .05 level, we multiply the standard error of the difference (4.95) by 1.96. The result is 9.70, or approximately 10 points. Thus the difference between an individual's WAIS Verbal and Performance IQ should be at least 10 points to be significant at the .05 level.

RELIABILITY OF CRITERION-REFERENCED TESTS

It will be recalled from Chapter 4 that criterion-referenced tests usually (but not necessarily) evaluate performance in terms of mastery rather than degree of achievement. A major statistical implication of mastery testing is a reduction in variability of scores among persons. Theoretically, if everyone continues training until the skill is mastered, variability is reduced to zero. Not only is low variability a result of the way such tests are used; it is also built into the tests through the construction and choice of items, as will be shown in Chapter 8.

In an earlier section of this chapter, we saw that *any* correlation, including reliability coefficients, is affected by the variability of the group in which it is computed. As the variability of the sample decreases, so does the correlation coefficient. Obviously, then, it would be inappropriate to assess the reliability of most criterion-referenced tests by the usual procedures.[9] Under these conditions, even a highly stable and internally consistent test could yield a reliability coefficient near zero.

In the construction of criterion-referenced tests, two important questions are: (1) How many items must be used for reliable assessment of each of the specific instructional objectives covered by the test? (2) What proportion of items must be correct for the reliable establishment of mastery? In much current testing, these two questions have been answered by judgmental decisions. Efforts are under way, however, to develop appropriate statistical techniques that will provide objective, empirical answers (see, e.g., Ferguson & Novick, 1973; Glaser & Nitko, 1971; Hambleton & Novick, 1973; Livingston, 1972; Millman, 1974). A few examples will serve to illustrate the nature and scope of these efforts.

The two questions about number of items and cutoff score can be incorporated into a single hypothesis, amenable to testing within the framework of decision theory and sequential analysis (Glaser & Nitko, 1971; Lindgren & McElrath, 1969; Wald, 1947). Specifically, we wish to test the hypothesis that the examinee has achieved the required level of mastery in the content domain or instructional objective sampled by the test items. Sequential analysis consists in taking observations one at a time and deciding after each observation whether to: (1) accept the hypothesis; (2) reject the hypothesis; or (3) make additional observations. Thus the number of observations (in this case number of items) needed to reach a reliable conclusion is itself determined during the process of testing. Rather than being presented with a fixed, predetermined number of items, the examinee continues taking the test until a mastery or nonmastery decision is reached. At that point, testing is discontinued and the student is either directed to the next instructional level or returned to the nonmastered level for further study. With the computer facilities described in Chapter 4, such sequential decision procedures are feasible and can reduce total testing time while yielding reliable estimates of mastery (Glaser & Nitko, 1971).

---

[9] For fuller discussion of special statistical procedures required for the construction and evaluation of criterion-referenced tests, see Glaser and Nitko (1971), Hambleton and Novick (1973), Millman (1974), Popham and Husek (1969). A set of tables for determining the minimum number of items required for establishing mastery at specified levels is provided by Millman (1972, 1973).

Some investigators have been exploring the use of Bayesian estimation techniques, which lend themselves well to the kind of decisions required by mastery testing. Because of the large number of specific instructional objectives to be tested, criterion-referenced tests typically provide only a small number of items for each objective. To supplement this limited information, procedures have been developed for incorporating collateral data from the student's previous performance history, as well as from the test results of other students (Ferguson & Novick, 1973; Hambleton & Novick, 1973).

When flexible, individually tailored procedures are impracticable, more traditional techniques can be utilized to assess the reliability of a given test. For example, mastery decisions reached at a prerequisite instructional level can be checked against performance at the next instructional level. Is there a sizeable proportion of students who reached or exceeded the cutoff score on the mastery test at the lower level and failed to achieve mastery at the next level within a reasonable period of instructional time? Does an analysis of their difficulties suggest that they had not truly mastered the prerequisite skills? If so, these findings would strongly suggest that the mastery test was unreliable. Either the addition of more items or the establishment of a higher cutoff score would seem to be indicated. Another procedure for determining the reliability of a mastery test is to administer two parallel forms to the same individuals and note the percentage of persons for whom the same decision (mastery or nonmastery) is reached on both forms (Hambleton & Novick, 1973).

In the development of several criterion-referenced tests, Educational Testing Service has followed an empirical procedure to set standards of mastery. This procedure involves administering the test in classes one grade below and one grade above the grade where the particular concept or skill is taught. The dichotomization can be further refined by using teacher judgments to exclude any cases in the lower grade known to have mastered the concept or skill and any cases in the higher grade who have demonstrably failed to master it. A cutting score, in terms of number or percentage of correct items, is then selected that best discriminates between the two groups.

All statistical procedures for use with criterion-referenced tests are in an exploratory stage. Much remains to be done, in both theoretical development and empirical tryouts, before the most effective methodology for different testing situations can be formulated.

CHAPTER 6

Validity: Basic Concepts

THE VALIDITY of a test concerns *what* the test measures and *how well* it does so. In this connection, we should guard against accepting the test name as an index of what the test measures. Test names provide short, convenient labels for identification purposes. Most test names are far too broad and vague to furnish meaningful clues to the behavior area covered, although increasing efforts are being made to use more specific and operationally definable test names. The trait measured by a given test can be defined only through an examination of the objective sources of information and empirical operations utilized in establishing its validity (Anastasi, 1950). Moreover, the validity of a test cannot be reported in general terms. No test can be said to have "high" or "low" validity in the abstract. Its validity must be determined with reference to the particular use for which the test is being considered.

Fundamentally, all procedures for determining test validity are concerned with the relationships between performance on the test and other independently observable facts about the behavior characteristics under consideration. The specific methods employed for investigating these relationships are numerous and have been described by various names. In the *Standards for Educational and Psychological Tests* (1974), these procedures are classified under three principal categories: content, criterion-related, and construct validity. Each of these types of validation procedures will be considered in one of the following sections, and the relations among them will be examined in a concluding section. Techniques for analyzing and interpreting validity data with reference to practical decisions will be discussed in Chapter 7.

CONTENT VALIDITY

NATURE. Content validity involves essentially the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured. Such a validation procedure is commonly used in evaluating achievement tests. This type of test is designed to measure how well the individual has mastered a specific skill or course of study. It might thus appear that mere inspection of the content of the test should suffice to establish its validity for such a purpose. A test of multiplication, spelling, or bookkeeping would seem to be valid by definition if it consists of multiplication, spelling, or bookkeeping items, respectively.

The solution, however, is not so simple as it appears to be.[1] One difficulty is that of adequately sampling the item universe. The behavior domain to be tested must be systematically analyzed to make certain that all major aspects are covered by the test items, and in the correct proportions. For example, a test can easily become overloaded with those aspects of the field that lend themselves more readily to the preparation of objective items. The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared. A well-constructed achievement test should cover the objectives of instruction, not just its subject matter. Content must therefore be broadly defined to include major objectives, such as the application of principles and the interpretation of data, as well as factual knowledge. Moreover, content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content. Mere inspection of the test may fail to reveal the processes actually used by examinees in taking the test.

It is also important to guard against any tendency to overgeneralize regarding the domain sampled by the test. For instance, a multiple-choice spelling test may measure the ability to recognize correctly and incorrectly spelled words. But it cannot be assumed that such a test also measures ability to spell correctly from dictation, frequency of misspellings in written compositions, and other aspects of spelling ability (Ahlström, 1964; Knoell & Harris, 1952). Still another difficulty arises from the possible inclusion of irrelevant factors in the test scores. For example, a test designed to measure proficiency in such areas as mathematics or mechanics may be unduly influenced by the ability to understand verbal directions or by speed of performing simple, routine tasks.

SPECIFIC PROCEDURES. Content validity is built into a test from the outset through the choice of appropriate items. For educational tests, the preparation of items is preceded by a thorough and systematic examination of relevant course syllabi and textbooks, as well as by consultation with subject-matter experts. On the basis of the information thus gathered, *test specifications* are drawn up for the item writers. These specifications should show the content areas or topics to be covered, the instructional objectives or processes to be tested, and the relative importance of individual topics and processes. On this basis, the number of items of each kind to be prepared on each topic can be established. A convenient way to set up such specifications is in terms of a two-way table, with processes across the top and topics in the left-hand column (see Table 30, Ch. 14). Not all cells in such a table, of course, need to have items, since certain processes may be unsuitable or irrelevant for certain topics. It might be added that such a specification table will also prove helpful in the preparation of teacher-made examinations for classroom use in any subject.

---

[1] Further discussions of content validity from several angles can be found in Ebel (1956), Huddleston (1956), and Lennon (1956).

In listing objectives to be covered in an educational achievement test, the test constructor can be guided by the extensive survey of educational objectives given in the *Taxonomy of Educational Objectives* (Bloom et al., 1956; Krathwohl et al., 1964). Prepared by a group of specialists in educational measurement, this handbook also provides examples of many types of items designed to test each objective. Two volumes are available, covering cognitive and affective domains, respectively. The major categories given in the cognitive domain include knowledge (in the sense of remembered facts, terms, methods, principles, etc.), comprehension, application, analysis, synthesis, and evaluation. The classification of affective objectives, concerned with the modification of attitudes, interests, values, and appreciation, includes five major categories: receiving, responding, valuing, organization, and characterization.

The discussion of content validity in the manual of an achievement test should include information on the content areas and the skills or objectives covered by the test, with some indication of the number of items in each category. In addition, the procedures followed in selecting categories and classifying items should be described. If subject-matter experts participated in the test-construction process, their number and professional qualifications should be stated. If they served as judges in classifying items, the directions they were given should be reported, as well as the extent of agreement among judges. Because curricula and course content change over time, it is particularly desirable to give the dates when subject-matter experts were consulted. Information should likewise be provided about number and nature of course syllabi and textbooks surveyed, including publication dates.

A number of empirical procedures may also be followed in order to supplement the content validation of an achievement test. Both total scores and performance on individual items can be checked for grade progress. In general, those items are retained that show the largest gains in the percentages of children passing them from the lower to the upper

FIG. 14. Data on Content Validity, as Illustrated by a Portion of Item Classification Table from STEP—Series II Reading Test, Level 3, for Grades 7 to 9.

(From Teacher's Manual for STEP—Series II. Copyright © 1971 by Educational Testing Service. All rights reserved. Reprinted by permission.)

Item Number	Percent Passing (Norms)			Comprehension				Translation and Inference				Analysis				Item Number
	Grade 7 % Right	Grade 8 % Right	Grade 9 % Right	Narrative	Humanities	Science	Social Studies	Narrative	Humanities	Science	Social Studies	Narrative	Humanities	Science	Social Studies	
1	77	83	88			x										1
2	61	71	82							x						2
3	49	53	60							x						3
4	32	36	45			x										4
5	54	64	73			x										5
6	35	41	49							x						6
7	59	67	74										x			7
8	24	28	32		x											8
9	41	46	51										x			9
10	68	74	82		x											10
11	40	47	47		x											11
12	53	63	66										x			12
13	59	68	73			x										13
14	60	70	76							x						14
15	58	69	77							x						15
16	21	22	27			x										16
17	41	50	61							x						17
18	37	42	48			x										18
19	28	34	47								x					19
20	31	39	45								x					20
21	45	50	58				x									21
22	32	38	49								x					22
23	65	76	82									x				23
24	54	64	69	x												24
25	48	50	52					x								25
26	19	27	28									x				26
27	22	26	36									x				27
28	25	32	43					x								28
29	29	36	42	x												29
30	24	30	40					x								30

grades. Figure 14 shows a portion of a table from the manual of the Sequential Tests of Educational Progress—Series II (STEP). For every item in each test in this achievement battery, the information provided includes its classification with regard to learning skill and type of material, as well as the percentage of children in the normative sample who gave the right answer to the item in each of the grades for which that level of the test is designed. The 30 items included in Figure 14 represent one part of the Reading test for Level 3, which covers grades 7 to 9.

Other supplementary procedures that may be employed, when appropriate, include analyses of types of errors commonly made on a test and observation of the work methods employed by examinees. The latter could be done by testing students individually with instructions to "think aloud" while solving each problem. The contribution of speed can be checked by noting how many persons fail to finish the test or by one of the more refined methods discussed in Chapter 5. To detect the possible irrelevant influence of ability to read instructions on test performance, scores on the test can be correlated with scores on a reading comprehension test. On the other hand, if the test is designed to measure reading comprehension, giving the questions without the reading passage on which they are based will show how many could be answered simply from the examinees' prior information or other irrelevant cues.

APPLICATIONS. Especially when bolstered by such empirical checks as those illustrated above, content validity provides an adequate technique for evaluating achievement tests. It permits us to answer two questions that are basic to the validity of an achievement test: (1) Does the test cover a representative sample of the specified skills and knowledge? (2) Is test performance reasonably free from the influence of irrelevant variables?

Content validity is particularly appropriate for the criterion-referenced tests described in Chapter 4. Because performance on these tests is interpreted in terms of content meaning, it is obvious that content validity is a prime requirement for their effective use. Content validation is also applicable to certain occupational tests designed for employee selection and classification, to be discussed in Chapter 15. This type of validation is suitable when the test is an actual job sample or otherwise calls for the same skills and knowledge required on the job. In such cases, a thorough job analysis should be carried out in order to demonstrate the close resemblance between the job activities and the test.

For aptitude and personality tests, on the other hand, content validity is usually inappropriate and may, in fact, be misleading. Although considerations of relevance and effectiveness of content must obviously enter into the initial stages of constructing any test, eventual validation of aptitude or personality tests requires empirical verification by the procedures to be described in the following sections. These tests bear less intrinsic resemblance to the behavior domain they are trying to sample than do achievement tests. Consequently, the content of aptitude and personality tests can do little more than reveal the hypotheses that led the test constructor to choose a certain type of content for measuring a specified trait. Such hypotheses need to be empirically confirmed to establish the validity of the test.

Unlike achievement tests, aptitude and personality tests are not based on a specified course of instruction or uniform set of prior experiences from which test content can be drawn. Hence, in the latter tests, individuals are likely to vary more in the work methods or psychological processes employed in responding to the same test items. The identical test might thus measure different functions in different persons. Under these conditions, it would be virtually impossible to determine the psychological functions measured by the test from an inspection of its content. For example, college graduates might solve a problem in verbal or mathematical terms, while a mechanic would arrive at the same solution in terms of spatial visualization. Or a test measuring arithmetic reasoning among high school freshmen might measure only individual differences in speed of computation when given to college students. A specific illustration of the dangers of relying on content analysis of aptitude tests is provided by a study conducted with a digit-symbol substitution test (Burik, 1950). This test, generally regarded as a typical "code-learning" test, was found to measure chiefly motor speed in a group of high school students.

FACE VALIDITY. Content validity should not be confused with face validity. The latter is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test "looks valid" to the examinees who take it, the administrative personnel who decide on its use, and other technically untrained observers. Fundamentally, the question of face validity concerns rapport and public relations. Although common usage of the term validity in this connection may make for confusion, face validity itself is a desirable feature of tests. For example, when tests originally designed for children and developed within a classroom setting were first extended for adult use, they frequently met with resistance and criticism because of their lack of face validity. Certainly if test content appears irrelevant, inappropriate, silly, or childish, the result will be poor cooperation, regardless of the actual validity of the test. Especially in adult testing, it is not sufficient for a test to be objectively valid. It also needs face validity to function effectively in practical situations.

Face validity can often be improved by merely reformulating test items in terms that appear relevant and plausible in the particular setting in which they will be used. For example, if a test of simple arithmetic reasoning is constructed for use with machinists, the items should be worded in terms of machine operations rather than in terms of "how many oranges can be purchased for 36 cents" or other traditional schoolbook problems. Similarly, an arithmetic test for naval personnel can be expressed in naval terminology, without necessarily altering the functions measured. To be sure, face validity should never be regarded as a substitute for objectively determined validity. It cannot be assumed that improving the face validity of a test will improve its objective validity. Nor can it be assumed that when a test is modified so as to increase its face validity, its objective validity remains unaltered. The validity of the test in its final form will always need to be directly checked.

CRITERION-RELATED VALIDITY

Criterion-related validity indicates the effectiveness of a test in predicting an individual's behavior in specified situations. For this purpose, performance on the test is checked against a *criterion,* i.e., a direct and independent measure of that which the test is designed to predict. Thus, for a mechanical aptitude test, the criterion might be subsequent job performance as a machinist; for a scholastic aptitude test, it might be college grades; and for a neuroticism test, it might be associates' ratings or other available information on the subjects' behavior in various life situations.

CONCURRENT AND PREDICTIVE VALIDITY. The criterion measure against which test scores are validated may be obtained at approximately the same time as the test scores or after a stated interval. The APA test *Standards* (1974), differentiate between concurrent and predictive validity on the basis of these time relations between criterion and test. The term "prediction" can be used in the broader sense, to refer to prediction from the test to any criterion situation, or in the more limited sense of prediction over a time interval. It is in the latter sense that it is used in the expression "predictive validity." The information provided by predictive validity is most relevant to tests used in the selection and classification of personnel. Hiring job applicants, selecting students for admission to college or professional schools, and assigning military personnel to occupational training programs represent examples of the sort of decisions requiring a knowledge of the predictive validity of tests. Other examples include the use of tests to screen out applicants likely to develop emotional disorders in stressful environments and the use of tests to identify psychiatric patients most likely to benefit from a particular therapy.

In a number of instances, concurrent validity is found merely as a substitute for predictive validity. It is frequently impracticable to extend validation procedures over the time required for predictive validity or to obtain a suitable preselection sample for testing purposes. As a compromise solution, therefore, tests are administered to a group on whom criterion data are already available. Thus, the test scores of college students may be compared with their cumulative grade-point average at the time of testing, or those of employees compared with their current job success.

For certain uses of psychological tests, on the other hand, concurrent validity is the most appropriate type and can be justified in its own right. The logical distinction between predictive and concurrent validity is based, not on time, but on the objectives of testing. Concurrent validity is relevant to tests employed for *diagnosis* of existing status, rather than prediction of future outcomes. The difference can be illustrated by asking: "Is Smith neurotic?" (concurrent validity) and "Is Smith likely to become neurotic?" (predictive validity).

Because the criterion for concurrent validity is always available at the time of testing, we might ask what function is served by the test in such situations. Basically, such tests provide a simpler, quicker, or less expensive substitute for the criterion data. For example, if the criterion consists of continuous observation of a patient during a two-week hospitalization period, a test that could sort out normals from neurotic and doubtful cases would appreciably reduce the number of persons requiring such extensive observation.

CRITERION CONTAMINATION. An essential precaution in finding the validity of a test is to make certain that the test scores do not themselves influence any individual's criterion status. For example, if a college instructor or a foreman in an industrial plant knows that a particular individual scored very poorly on an aptitude test, such knowledge might influence the grade given to the student or the rating assigned to the worker. Or a high-scoring person might be given the benefit of the doubt when academic grades or on-the-job ratings are being prepared. Such influences would obviously raise the correlation between test scores and criterion in a manner that is entirely spurious or artificial.

This possible source of error in test validation is known as criterion contamination, since the criterion ratings become "contaminated" by the rater's knowledge of the test scores. To prevent the operation of such an error, it is absolutely essential that no person who participates in the assignment of criterion ratings have any knowledge of the examinees' test scores. For this reason, test scores employed in "testing the test" must be kept strictly confidential. It is sometimes difficult to convince teachers, employers, military officers, and other line personnel that such a precaution is essential. In their urgency to utilize all available information for practical decisions, such persons may fail to realize that the test scores must be put aside until the criterion data mature and validity can be checked.

COMMON CRITERIA. Any test may be validated against as many criteria as there are specific uses for it. Any method for assessing behavior in any situation could provide a criterion measure for some particular purpose. The criteria employed in finding the validities reported in test manuals, however, fall into a few common categories. Among the criteria most frequently employed in validating intelligence tests is some index of *academic achievement*. It is for this reason that such tests have often been more precisely described as measures of scholastic aptitude. The specific indices used as criterion measures include school grades, achievement test scores, promotion and graduation records, special honors and awards, and teachers' or instructors' ratings for "intelligence." Insofar as such ratings given within an academic setting are likely to be heavily colored by the individual's scholastic performance, they may be properly classified with the criterion of academic achievement.

The various indices of academic achievement have provided criterion data at all educational levels, from the primary grades to college and graduate school. Although employed principally in the validation of general intelligence tests, they have also served as criteria for certain multiple-aptitude and personality tests. In the validation of any of these types of tests for use in the selection of college students, for example, a common criterion is freshman grade-point average. This measure is the average grade in all courses taken during the freshman year, each grade being weighted by the number of course points for which it was received.

A variant of the criterion of academic achievement frequently employed with out-of-school adults is the amount of education the individual completed. It is expected that in general the more intelligent individuals continue their education longer, while the less intelligent drop out of school earlier. The assumption underlying this criterion is that the educational ladder serves as a progressively selective influence, eliminating those incapable of continuing beyond each step. Although it is undoubtedly true that college graduates, for example, represent a more highly selected group than elementary school graduates, the relation between amount of education and scholastic aptitude is far from perfect. Especially at the higher educational levels, economic, social, motivational, and other nonintellectual factors may influence the continuation of the individual's education. Moreover, with such concurrent validation, it is difficult to disentangle cause-and-effect relations. To what extent are the obtained differences in intelligence test scores simply the result of the varying amount of education? And to what extent could the test have predicted individual differences in subsequent educational progress? These questions can be answered only when the test is administered before the criterion data have matured, as in predictive validation.

In the development of special aptitude tests, a frequent type of criterion is based on *performance in specialized training.* For example, mechanical aptitude tests may be validated against final achievement in shop courses. Various business school courses, such as stenography, typing, or bookkeeping, provide criteria for aptitude tests in these areas. Similarly, performance in music or art schools has been employed in validating music or art aptitude tests. Several professional aptitude tests have been validated in terms of achievement in schools of law, medicine, dentistry, engineering, and other areas. In the case of custom-made tests, designed for use within a specific testing program, training records are a frequent source of criterion data. An outstanding illustration is the validation of Air Force pilot selection tests against performance in basic flight training. Performance in training programs is also commonly used as a criterion for test validation in other military occupational specialties and in some industrial validation studies.

Among the specific indices of training performance employed for criterion purposes may be mentioned achievement tests administered on completion of training, formally assigned grades, instructors' ratings, and successful completion of training versus elimination from the program. Multiple aptitude batteries have often been checked against grades in specific high school or college courses, in order to determine their validity as differential predictors. For example, scores on a verbal comprehension test may be compared with grades in English courses, spatial visualization scores with geometry grades, and so forth.

In connection with the use of training records in general as criterion measures, a useful distinction is that between intermediate and ultimate criteria. In the development of an Air Force pilot-selection test or a medical aptitude test, for example, the ultimate criteria would be combat performance and eventual achievement as a practicing physician, respectively. Obviously it would require a long time for such criterion data to mature. It is doubtful, moreover, whether a truly ultimate criterion is ever obtained in actual practice. Finally, even were such an ultimate criterion available, it would probably be subject to many uncontrolled factors that would render it relatively useless. For example, it would be difficult to evaluate the relative degree of success of physicians practicing different specialties and in different parts of the country. For these reasons, such intermediate criteria as performance records at some stage of training are frequently employed as criterion measures.

For many purposes, the most satisfactory type of criterion measure is that based on follow-up records of actual *job performance*. This criterion has been used to some extent in the validation of general intelligence as well as personality tests, and to a larger extent in the validation of special aptitude tests. It is a common criterion in the validation of custom-made tests for specific jobs. The "jobs" in question may vary widely in both level and kind, including work in business, industry, the professions, and the armed services. Most measures of job performance, although probably not representing ultimate criteria, at least provide good intermediate criteria for many testing purposes. In this respect they are to be preferred to training records. On the other hand, the measurement of job performance does not permit as much uniformity of conditions as is possible during training. Moreover, since it usually involves a longer follow-up, the criterion of job performance is likely to entail a loss in the number of available subjects. Because of the variation in the nature of nominally similar jobs in different organizations, test manuals reporting validity data against job criteria should describe not only the specific criterion measures employed but also the job duties performed by the workers.

Validation by the method of *contrasted groups* generally involves a composite criterion that reflects the cumulative and uncontrolled selective influences of everyday life. This criterion is ultimately based on survival within a particular group versus elimination therefrom. For example, in the validation of an intelligence test, the scores obtained by institutionalized mentally retarded children may be compared with those obtained by schoolchildren of the same age. In this case, the multiplicity of factors determining commitment to an institution for the mentally retarded constitutes the criterion. Similarly, the validity of a musical aptitude or a mechanical aptitude test may be checked by comparing the scores obtained by students enrolled in a music school or an engineering school, respectively, with the scores of unselected high school or college students.

To be sure, contrasted groups can be selected on the basis of any criterion, such as school grades, ratings, or job performance, by simply choosing the extremes of the distribution of criterion measures. The contrasted groups included in the present category, however, are distinct groups that have gradually become differentiated through the operation of the multiple demands of daily living. The criterion under consideration is thus more complex and less clearly definable than those previously discussed.

The method of contrasted groups is used quite commonly in the validation of personality tests. Thus, in validating a test of social traits, the test performance of salesmen or executives, on the one hand, may be compared with that of clerks or engineers, on the other. The assumption underlying such a procedure is that, with reference to many social traits, individuals who have entered and remained in such occupations as selling or executive work will as a group excel persons in such fields as clerical work or engineering. Similarly, college students who have engaged in many extracurricular activities may be compared with those who have participated in none during a comparable period of college attendance. Occupational groups have frequently been used in the development and validation of interest tests, such as the Strong Vocational Interest Blank, as well as in the preparation of attitude scales. Other groups sometimes employed in the validation of attitude scales include political, religious, geographical, or other special groups generally known to represent distinctly different points of view on certain issues.

In the development of certain personality tests, *psychiatric diagnosis* is used both as a basis for the selection of items and as evidence of test validity. Psychiatric diagnosis may serve as a satisfactory criterion provided that it is based on prolonged observation and detailed case history, rather than on a cursory psychiatric interview or examination. In the latter case, there is no reason to expect the psychiatric diagnosis to be superior to the test score itself as an indication of the individual's emotional condition. Such a psychiatric diagnosis could not be regarded as a criterion measure, but rather as an indicator or predictor whose own validity would have to be determined.

Mention has already been made, in connection with other criterion categories, of certain types of *ratings* by school teachers, instructors in specialized courses, and job supervisors. To these can be added ratings by officers in military situations, ratings of students by school counselors, and ratings by co-workers, classmates, fellow club-members, and other groups of associates. The ratings discussed earlier represented merely a subsidiary technique for obtaining information regarding such criteria as academic achievement, performance in specialized training, or job success. We are now considering the use of ratings as the very core of the criterion measure. Under these circumstances, the ratings themselves define the criterion. Moreover, such ratings are not restricted to the evaluation of specific achievement, but involve a personal judgment by an observer regarding any of the variety of traits that psychological tests attempt to measure. Thus, the subjects in the validation sample might be rated on such characteristics as dominance, mechanical ingenuity, originality, leadership, or honesty.

Ratings have been employed in the validation of almost every type of test. They are particularly useful in providing criteria for personality tests, since objective criteria are much more difficult to find in this area. This is especially true of distinctly social traits, in which ratings based on personal contact may constitute the most logically defensible criterion. Although ratings may be subject to many judgmental errors, when obtained under carefully controlled conditions they represent a valuable source of criterion data. Techniques for improving the accuracy of ratings and for reducing common types of errors will be considered in Chapter 20.

Finally, correlations between a new test and *previously available tests* are frequently cited as evidence of validity. When the new test is an abbreviated or simplified form of a currently available test, the latter can properly be regarded as a criterion measure. Thus, a paper-and-pencil test might be validated against a more elaborate and time-consuming performance test whose validity had previously been established. Or a group test might be validated against an individual test. The Stanford-Binet, for example, has repeatedly served as a criterion in validating group tests. In such a case, the new test may be regarded at best as a crude approximation of the earlier one. It should be noted that unless the new test represents a simpler or shorter substitute for the earlier test, the use of the latter as a criterion is indefensible.

SPECIFICITY OF CRITERIA. Criterion-related validity is most appropriate for local validation studies, in which the effectiveness of a test for a specific program is to be assessed. This is the approach followed, for example, when a given company wishes to evaluate a test for selecting applicants for one of its jobs or when a given college wishes to determine how well an academic aptitude test can predict the course performance of its students. Criterion-related validity can be best characterized as the practical validity of a test in a specified situation. This type of validation represents applied research, as distinguished from basic research, and as such it provides results that are less generalizable than the results of other procedures.

That criterion-related validity may be quite specific has been demonstrated repeatedly. Figure 15 gives examples of the wide variation in the correlations of a single type of test with criteria of job proficiency. The first graph shows the distribution of 72 correlations found between intelligence test scores and measures of the job proficiency of general clerks; the second graph summarizes in similar fashion 191 correlations between finger dexterity tests and the job proficiency of benchworkers. Although in both instances the correlations tend to cluster in a particular range of validity, the variation among individual studies is considerable. The validity coefficient may be high and positive in one study and negligible or even substantially negative in another.

Similar variation with regard to the prediction of course grades is illustrated in Figure 16. This figure shows the distribution of correlations obtained between grades in mathematics and scores on each of the subtests of the Differential Aptitude Tests. Thus, for the Numerical Ability test (NA), the largest number of validity coefficients among boys fell between .50 and .59; but the correlations obtained in different mathematics courses and in different schools ranged from .22 to .75. Equally wide differences were found with the other subtests and, it might be added, with grades in other subjects not included in Figure 16.

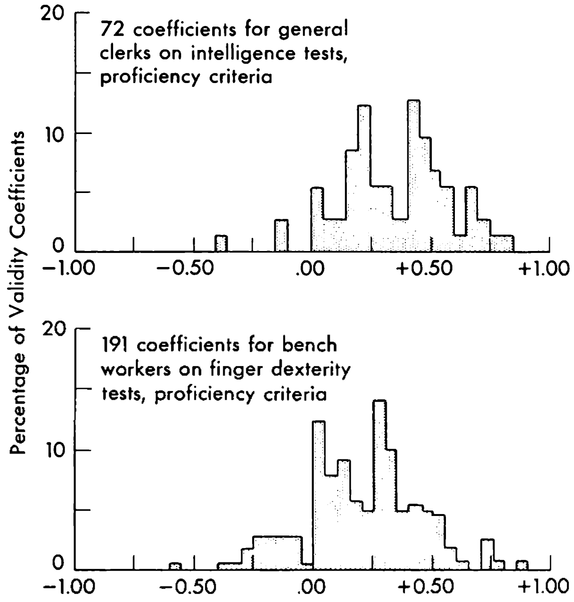
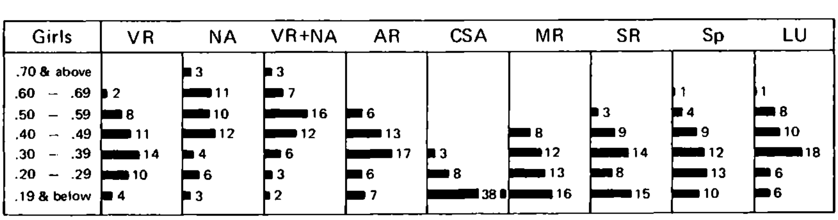


FIG. 15. Examples of Variation in Validity Coefficients of Given Tests for Particular Jobs.

(Adapted from Ghiselli, 1966, p. 29.)

Some of the variation in validity coefficients against job criteria reported in Figure 15 results from differences among the specific tests employed in different studies to measure intelligence or finger dexterity. In the results of both Figures 15 and 16, moreover, some variation is attributable to differences in homogeneity and level of the groups tested. The range of validity coefficients found, however, is far wider than could be explained in these terms. Differences in the criteria themselves are undoubtedly a major reason for the variation observed among validity coefficients. Thus, the duties of office clerks or benchworkers may differ widely among companies or among departments in the same company. Similarly, courses in the same subject may differ in content, teaching method, instructor characteristics, bases for evaluating student achievement, and numerous other ways. Consequently, what appears to be the same criterion may represent a very different combination of traits in different situations.

Criteria may also vary over time in the same situation. For example, the validity coefficient of a test against job training criteria often differs from its validity against job performance criteria (Ghiselli, 1966). There is evidence that the traits required for successful performance of a given job or even a single task vary with the amount of practice or job experience of the individual (Fleishman, 1972; Fleishman & Fruchter, 1960; Fleishman & Hempel, 1956; Ghiselli & Haire, 1960). There is also abundant evidence to show that job criteria change over time for other reasons, such as changing nature of jobs, shifts in organizational goals, individual advancement in rank, and other temporal conditions (MacKinney, 1967; Prien, 1966). It is well known, of course, that educational curricula and course content change over time. In other words, the criteria most commonly used in validating intelligence and aptitude tests—namely, job performance and educational achievement—are dynamic



Mathematics

Boys	VR	NA	VR+NA	AR	CSA	MR	SR	Sp	LU
.70 & above		3	1						
.60 – .69		8	8	2		1	1		
.50 – .59	6	19	13	6		2	3	3	7
.40 – .49	8	8	12	10		2	7	7	11
.30 – .39	14	9	12	13	2	6	12	7	16
.20 – .29	12	4	4	10	10	18	10	12	9
.19 & below	11		1	10	39	22	18	22	8

Girls	VR	NA	VR+NA	AR	CSA	MR	SR	Sp	LU
.70 & above		3	3						
.60 – .69	2	11	7					1	1
.50 – .59	8	10	16	6			3	4	8
.40 – .49	11	12	12	13		8	9	9	10
.30 – .39	14	4	6	17	3	12	14	12	18
.20 – .29	10	6	3	6	8	13	8	13	6
.19 & below	4	3	2	7	38	16	15	10	6

Fig. 16. Graphic Summary of Validity Coefficients of the Differential Aptitude Tests (Forms S and T) for Course Grades in Mathematics. The bars and accompanying numbers in each column indicate the number of coefficients in the ranges given at the left.

(From Fifth Edition Manual, p. 82. Reproduced by permission. Copyright © 1973, 1974 by The Psychological Corporation, New York, N.Y. All rights reserved.)

rather than static. It follows that criterion-related validity is itself subject to temporal changes.

SYNTHETIC VALIDITY. Criteria not only differ across situations and over time, but they are also likely to be complex (see, e.g., Richards, Taylor, Price, & Jacobsen, 1965). Success on a job, in school, or in other activities of daily life depends not on one trait but on many traits. Hence, practical criteria are likely to be multifaceted. Several different indicators or measures of job proficiency or academic achievement could thus be used in validating a test. Since these measures may tap different traits or combinations of traits, it is not surprising to find that they yield different validity coefficients for any given test. When different criterion measures are obtained for the same individuals, their intercorrelations are often quite low. For instance, accident records or absenteeism may show virtually no relation to productivity or error data for the same job (Seashore, Indik, & Georgopoulos, 1960). These differences, of course, are reflected in the validity coefficients of any given test against different criterion measures. Thus, a test may fail to correlate significantly with supervisors' ratings of job proficiency and yet show appreciable validity in predicting who will resign and who will be promoted at a later date (Albright, Smith, & Glennon, 1959).

Because of criterion complexity, validating a test against a composite criterion of job proficiency, academic achievement, or other similar accomplishments may be of questionable value and is certainly of limited generality. If different subcriteria are relatively independent, a more effective procedure is to validate each test against that aspect of the criterion it is best designed to measure. An analysis of these more specific relationships lends *meaning* to the test scores in terms of the multiple dimensions of criterion behavior (Dunnette, 1963; Ebel, 1961; S. R. Wallace, 1965). For example, one test might prove to be a valid predictor of a clerk's perceptual speed and accuracy in handling detail work, another of his ability to spell correctly, and still another of his ability to resist distraction.

If, now, we return to the practical question of evaluating a test or combination of tests for effectiveness in predicting a complex criterion such as success on a given job, we are faced with the necessity of conducting a separate validation study in each local situation and repeating it at frequent intervals. This is admittedly a desirable procedure and one that is often recommended in test manuals. In many situations, however, it is not feasible to follow this procedure because of well-nigh insurmountable practical obstacles. Even if adequately trained personnel are available to carry out the necessary research, most criterion-related validity studies conducted in industry are likely to prove unsatisfactory for at least three reasons. First, it is difficult to obtain dependable and sufficiently comprehensive criterion data. Second, the number of employees engaged in the same or closely similar jobs within a company is often too small for significant statistical results. Third, correlations will very probably be lowered by restriction of range through preselection, since only those persons actually hired can be followed up on the job.

For all the reasons discussed above, personnel psychologists have shown increasing interest in a technique known as synthetic validity. First introduced by Lawshe (1952), the concept of synthetic validity has been defined by Balma (1959, p. 395) as "the inferring of validity in a specific situation from a systematic analysis of job elements, a determination of test validity for these elements, and a combination of elemental validities into a whole." Several procedures have been developed for gathering the needed empirical data and for combining these data to obtain an estimate of synthetic validity for a particular complex criterion (see, e.g., Guion, 1965; Lawshe & Balma, 1966, Ch. 14; McCormick, 1959; Primoff, 1959, 1975). Essentially, the process involves three steps: (1) detailed job analysis to identify the job elements and their relative weights; (2) analysis and empirical study of each test to determine the extent to which it measures proficiency in performing each of these job elements; and (3) finding the validity of each test for the given job synthetically from the weights of these elements in the job and in the test.

In a long-term research program conducted with U.S. Civil Service job applicants, Primoff (1975) has developed the J-coefficient (for "job-coefficient") as an index of synthetic validity. Among the special features of this procedure are the listing of job elements expressed in terms of worker behavior and the rating of the relative importance of these elements in each job by supervisors and job incumbents. Correlations between test scores and self-ratings on job elements are found in total applicant samples (not subject to the preselection of employed workers). Various checking procedures are followed to ensure stability of correlations and weights derived from self-ratings, as well as adequacy of criterion coverage. For these purposes, data are obtained from different samples of applicant populations. The final estimate of correlation between test and job performance is found from the correlation of each job element with the particular job and the weight of the same element in the given test.[2] There is evidence that the J-coefficient has proved helpful in improving the employment opportunities of minority applicants and persons with little formal education, because of its concentration on job-relevant skills (Primoff, 1975).

---

[2] The statistical procedures are essentially an adaptation of multiple regression equations, to be discussed in Chapter 7. For each job element, its correlation with the job is multiplied by its weight in the test, and these products are added across all appropriate job elements.

A different application of synthetic validity, especially suitable for use in a small company with few employees in each type of job, is described by Guion (1965). The study was carried out in a company having 48 employees, each of whom was doing a job that was appreciably different from the jobs of the other employees. Detailed job analyses nevertheless revealed seven job elements common to many jobs. Each employee was rated on the job elements appropriate to his job; and these ratings were then checked against the employees' scores on each test in a trial battery. On the basis of these analyses, a separate battery could be "synthesized" for each job by combining the two best tests for each of the job elements demanded by that job. When the batteries thus assembled were applied to a subsequently hired group of 13 employees, the results showed considerable promise. Because of the small number of cases, these results are only suggestive. The study was conducted primarily to demonstrate a model for the utilization of synthetic validity.

The two examples of synthetic validity were cited only to illustrate the scope of possible applications of these techniques. For a description of the actual procedures followed, the reader is referred to the original sources. In summary, the concept of synthetic validity can be implemented in different ways to fit the practical exigencies of different situations. It offers a promising approach to the problem of complex and changing criteria; and it permits the assembling of test batteries to fit the requirements of specific jobs and the determination of test validity in many contexts where adequate criterion-related validation studies are impracticable.

CONSTRUCT VALIDITY

The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait. Examples of such constructs are intelligence, mechanical comprehension, verbal fluency, speed of walking, neuroticism, and anxiety. Focusing on a broader, more enduring, and more abstract kind of behavioral description than the previously discussed types of validity, construct validation requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations are grist for this validity mill. Illustrations of specific techniques suitable for construct validation will be considered below.

DEVELOPMENTAL CHANGES. A major criterion employed in the validation of a number of intelligence tests is *age differentiation.* Such tests as the Stanford-Binet and most preschool tests are checked against chronological age to determine whether the scores show a progressive increase with advancing age. Since abilities are expected to increase with age during childhood, it is argued that the test scores should likewise show such an increase, if the test is valid. The very concept of an age scale of intelligence, as initiated by Binet, is based on the assumption that "intelligence" increases with age, at least until maturity.

The criterion of age differentiation, of course, is inapplicable to any functions that do not exhibit clear-cut and consistent age changes. In the area of personality measurement, for example, it has found limited use. Moreover, it should be noted that, even when applicable, age differentiation is a necessary but not a sufficient condition for validity. Thus, if the test scores fail to improve with age, such a finding probably indicates that the test is not a valid measure of the abilities it was designed to sample. On the other hand, to prove that a test measures something that increases with age does not define the area covered by the test very precisely. A measure of height or weight would also show regular age increments, although it would obviously not be designated as an intelligence test.

A final point should be emphasized regarding the interpretation of the age criterion. A psychological test validated against such a criterion measures behavior characteristics that increase with age under the conditions existing in the type of environment in which the test was standardized. Because different cultures may stimulate and foster the development of dissimilar behavior characteristics, it cannot be assumed that the criterion of age differentiation is a universal one. Like all other criteria, it is circumscribed by the particular cultural setting in which it is derived.

Developmental analyses are also basic to the construct validation of the Piagetian ordinal scales cited in Chapter 4. A fundamental assumption of such scales is the *sequential patterning* of development, such that the attainment of earlier stages in concept development is prerequisite to the acquisition of later conceptual skills. There is thus an intrinsic hierarchy in the content of these scales. The construct validation of ordinal scales should therefore include empirical data on the sequential invariance of the successive steps. This involves checking the performance of children at different levels in the development of any tested concept, such as conservation or object permanence. Do children who demonstrate mastery of the concept at a given level also exhibit mastery at the lower levels? Insofar as criterion-referenced tests are also frequently designed according to a hierarchical pattern of learned skills, they, too, can utilize empirical evidence of hierarchical invariance in their validation.

CORRELATIONS WITH OTHER TESTS. Correlations between a new test and similar earlier tests are sometimes cited as evidence that the new test measures approximately the same general area of behavior as other tests designated by the same name, such as "intelligence tests" or "mechanical aptitude tests." Unlike the correlations found in criterion-related validity, these correlations should be moderately high, but not too high. If the new test correlates too highly with an already available test, without such added advantages as brevity or ease of administration, then the new test represents needless duplication.

Correlations with other tests are employed in still another way to demonstrate that the new test is relatively free from the influence of certain irrelevant factors. For example, a special aptitude test or a personality test should have a negligible correlation with tests of general intelligence or scholastic aptitude. Similarly, reading comprehension should not appreciably affect performance on such tests. Thus, correlations with tests of general intelligence, reading, or verbal comprehension are sometimes reported as indirect or negative evidence of validity. In these cases, high correlations would make the test suspect. Low correlations, however, would not in themselves insure validity. It will be noted that this use of correlations with other tests is similar to one of the supplementary techniques described under content validity.

FACTOR ANALYSIS. Of particular relevance to construct validity is factor analysis, a statistical procedure for the identification of psychological traits. Essentially, factor analysis is a refined technique for analyzing the interrelationships of behavior data. For example, if 20 tests have been given to 300 persons, the first step is to compute the correlations of each test with every other. An inspection of the resulting table of 190 correlations may itself reveal certain clusters among the tests, suggesting the location of common traits. Thus, if such tests as vocabulary, analogies, opposites, and sentence completion have high correlations with each other and low correlations with all other tests, we could tentatively infer the presence of a verbal comprehension factor. Because such an inspectional analysis of a correlation table is difficult and uncertain, however, more precise statistical techniques have been developed to locate the common factors required to account for the obtained correlations. These techniques of factor analysis will be examined further in Chapter 13, together with multiple aptitude tests developed by means of factor analysis.

In the process of factor analysis, the number of variables or categories in terms of which each individual's performance can be described is reduced from the number of original tests to a relatively small number of factors, or common traits. In the example cited above, five or six factors might suffice to account for the intercorrelations among the 20 tests. Each individual might thus be described in terms of his scores in the five or six factors, rather than in terms of the original 20 scores. A major purpose of factor analysis is to simplify the description of behavior by reducing the number of categories from an initial multiplicity of test variables to a few common factors, or traits.

After the factors have been identified, they can be utilized in describing the factorial composition of a test. Each test can thus be characterized in terms of the major factors determining its scores, together with the weight or loading of each factor and the correlation of the test with each factor. Such a correlation is known as the *factorial validity* of the test. Thus, if the verbal comprehension factor has a weight of .66 in a vocabulary test, the factorial validity of this vocabulary test as a measure of the trait of verbal comprehension is .66. It should be noted that factorial validity is essentially the correlation of the test with whatever is common to a group of tests or other indices of behavior. The set of variables analyzed can, of course, include both test and nontest data. Ratings and other criterion measures can thus be utilized, along with other tests, to explore the factorial validity of a particular test and to define the common traits it measures.

INTERNAL CONSISTENCY. In the published descriptions of certain tests, especially in the area of personality, the statement is made that the test has been validated by the method of internal consistency. The essential characteristic of this method is that the criterion is none other than the total score on the test itself. Sometimes an adaptation of the contrasted group method is used, extreme groups being selected on the basis of the total test score. The performance of the upper criterion group on each test item is then compared with that of the lower criterion group. Items that fail to show a significantly greater proportion of "passes" in the upper than in the lower criterion group are considered invalid, and are either eliminated or revised. Correlational procedures may also be employed for this purpose. For example, the biserial correlation between "pass-fail" on each item and total test score can be computed. Only those items yielding significant item-test correlations would be retained. A test whose items were selected by this method can be said to show internal consistency, since each item differentiates in the same direction as the entire test.

Another application of the criterion of internal consistency involves the correlation of subtest scores with total score. Many intelligence tests, for instance, consist of separately administered subtests (such as vocabulary, arithmetic, picture completion, etc.) whose scores are combined in finding the total test score. In the construction of such tests, the scores on each subtest are often correlated with total score and any subtest whose correlation with total score is too low is eliminated. The correlations of the remaining subtests with total score are then reported as evidence of the internal consistency of the entire instrument.

It is apparent that internal consistency correlations, whether based on items or subtests, are essentially measures of homogeneity. Because it helps to characterize the behavior domain or trait sampled by the test, the degree of homogeneity of a test has some relevance to its construct validity. Nevertheless, the contribution of internal consistency data to test validation is very limited. In the absence of data external to the test itself, little can be learned about what a test measures.

EFFECT OF EXPERIMENTAL VARIABLES ON TEST SCORES. A further source of data for construct validation is provided by experiments on the effect of selected variables on test scores. In checking the validity of a criterion-referenced test for use in an individualized instructional program, for example, one approach is through a comparison of pretest and posttest scores. The rationale of such a test calls for low scores on the pretest, administered before the relevant instruction, and high scores on the posttest. This relationship can also be checked for individual items in the test (Popham, 1971). Ideally, the largest proportion of examinees should fail an item on the pretest and pass it on the posttest. Items that are commonly failed on both tests are too difficult, and those passed on both tests too easy, for the purposes of such a test. If a sizeable proportion of examinees pass an item on the pretest and fail it on the posttest, there is obviously something wrong with the item, or the instruction, or both.

A test designed to measure anxiety-proneness can be administered to subjects who are subsequently put through a situation designed to arouse anxiety, such as taking an examination under distracting and stressful conditions. The initial anxiety test scores can then be correlated with physiological and other indices of anxiety expression during and after the examination. A different hypothesis regarding an anxiety test could be evaluated by administering the test before and after an anxiety-arousing experience and seeing whether test scores rise significantly on the retest. Positive findings from such an experiment would indicate that the test scores reflect current anxiety level. In a similar way, experiments can be designed to test any other hypothesis regarding the trait measured by a given test.

CONVERGENT AND DISCRIMINANT VALIDATION. In a thoughtful analysis of construct validation, D. T. Campbell (1960) points out that in order to demonstrate construct validity we must show not only that a test correlates highly with other variables with which it should theoretically correlate, but also that it does not correlate significantly with variables from which it should differ. In an earlier article, Campbell and Fiske (1959) described the former process as convergent validation and the latter as discriminant validation. Correlation of a mechanical aptitude test with subsequent grades in a shop course would be an example of convergent validation. For the same test, discriminant validity would be illustrated by a low and insignificant correlation with scores on a reading comprehension test, since reading ability is an irrelevant variable in a test designed to measure mechanical aptitude.

It will be recalled that the requirement of low correlation with irrelevant variables was discussed in connection with supplementary and precautionary procedures followed in content validation. Discriminant validation is also especially relevant to the validation of personality tests, in which irrelevant variables may affect scores in a variety of ways.

Campbell and Fiske (1959) proposed a systematic experimental design for the dual approach of convergent and discriminant validation, which they called the *multitrait-multimethod matrix.* Essentially, this procedure requires the assessment of two or more traits by two or more methods. A hypothetical example provided by Campbell and Fiske will serve to illustrate the procedure. Table 12 shows all possible correlations among the scores obtained when three traits are each measured by three methods. The three traits could represent three personality characteristics, such as (A) dominance, (B) sociability, and (C) achievement motivation. The three methods could be (1) a self-report inventory, (2) a projective technique, and (3) associates' ratings. Thus, A_1 would indicate dominance scores on the self-report inventory, A_2 dominance scores on the projective test, C_3 associates' ratings on achievement motivation, and so forth.

The hypothetical correlations given in Table 12 include reliability coefficients (in parentheses, along principal diagonal) and validity coefficients (in boldface, along three shorter diagonals). In these validity coefficients, the scores obtained for the same trait by different methods are correlated; each measure is thus being checked against other, independent measures of the same trait, as in the familiar validation procedure. The table also includes correlations between *different* traits measured by the *same* method (in solid triangles) and correlations between *different* traits measured by *different* methods (in broken triangles). For satisfactory construct validity, the validity coefficients should obviously be higher than the correlations between different traits measured by different methods; they should also be higher than the correlations be-

TABLE 12

A Hypothetical Multitrait-Multimethod Matrix

(From Campbell & Fiske, 1959, p. 82.)

		Method 1			Method 2			Method 3		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(.89)								
	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
Method 2	A_2	.57	.22	.09	(.93)					
	B_2	.22	.57	.10	.68	(.94)				
	C_2	.11	.11	.46	.59	.58	(.84)			
Method 3	A_3	.56	.22	.11	.67	.42	.33	(.94)		
	B_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

Note: Letters A, B, C refer to traits, subscripts 1, 2, 3 to methods. Validity coefficients (monotrait-heteromethod) are the three diagonal sets of boldface numbers; reliability coefficients (monotrait-monomethod) are the numbers in parentheses along principal diagonal. Solid triangles enclose heterotrait-monomethod correlations; broken triangles enclose heterotrait-heteromethod correlations.

tween different traits measured by the same method. For example, the correlation between dominance scores from a self-report inventory and dominance scores from a projective test should be higher than the correlation between dominance and sociability scores from a self-report inventory. If the latter correlation, representing common method variance, were high, it might indicate, for example, that a person's scores on this inventory are unduly affected by some irrelevant common factor such as ability to understand the questions or desire to make oneself appear in a favorable light on all traits.

Fiske (1973) has added still another set of correlations that should be checked, especially in the construct validation of personality tests. These correlations involve the same trait measured by the same method, but with a different test. For example, two investigators may each prepare a self-report inventory designed to assess endurance. Yet the endurance scores obtained with the two inventories may show quite different patterns of correlations with measures of other personality traits. Under these conditions, it cannot be concluded that both inventories measure the same personality construct of endurance.

It might be noted that within the framework of the multitrait-multimethod matrix, reliability represents agreement between two measures of the same trait obtained through maximally similar methods, such as parallel forms of the same test; validity represents agreement between two measures of the same trait obtained by maximally different methods, such as test scores and supervisor's ratings. Since similarity and difference of methods are matters of degree, theoretically reliability and validity can be regarded as falling along a single continuum. Ordinarily, however, the techniques actually employed to measure reliability and validity correspond to easily identifiable regions of this continuum.

OVERVIEW

We have considered several ways of asking, "How valid is this test?" To point up the distinctive features of the different types of validity, let us apply each in turn to a test consisting of 50 assorted arithmetic problems. Four ways in which this test might be employed, together with the type of validation procedure appropriate to each, are illustrated in Table 13. This example highlights the fact that the choice of validation procedure depends on the use to be made of the test scores. The same test, when employed for different purposes, should be validated in different ways. If an achievement test is used to predict subsequent performance

TABLE 13
Validation of a Single Arithmetic Test for Different Purposes

Testing Purpose	Illustrative Question	Type of Validity
Achievement test in elementary school arithmetic	How much has Dick learned in the past?	Content
Aptitude test to predict performance in high school mathematics	How well will Jim learn in the future?	Criterion-related: predictive
Technique for diagnosing learning disabilities	Does Bill's performance show specific disabilities?	Criterion-related: concurrent
Measure of logical reasoning	How can we describe Henry's psychological functioning?	Construct

at a higher educational level, as when selecting high school students for college admission, it needs to be evaluated against the criterion of subsequent college performance rather than in terms of its content validity.

The examples given in Table 13 focus on the differences among the various types of validation procedures. Further consideration of these procedures, however, shows that content, criterion-related, and construct validity do not correspond to distinct or logically coordinate categories. On the contrary, construct validity is a comprehensive concept, which includes the other types. All the specific techniques for establishing content and criterion-related validity, discussed in earlier sections of this chapter, could have been listed again under construct validity. Comparing the test performance of contrasted groups, such as neurotics and normals, is one way of checking the construct validity of a test designed to measure emotional adjustment, anxiety, or other postulated traits. Comparing the test scores of institutionalized mental retardates with those of normal schoolchildren is one way to investigate the construct validity of an intelligence test. The correlations of a mechanical aptitude test with performance in shop courses and in a wide variety of jobs contribute to our understanding of the construct measured by the test. Validity against various practical criteria is commonly reported in test manuals to aid the potential user in understanding what a test measures. Although he may not be directly concerned with the prediction of any of the specific criteria employed, by examining such criteria the test user is able to build up a concept of the behavior domain sampled by the test.

Content validity likewise enters into both the construction and the subsequent evaluation of all tests. In assembling items for any new test, the test constructor is guided by hypotheses regarding the relations between the type of content he chooses and the behavior he wishes to measure. All the techniques of criterion-related validation, as well as the other techniques discussed under construct validation, represent ways of testing such hypotheses. As for the test user, he too relies in part on content validity in evaluating any test. For example, he may check the vocabulary in an emotional adjustment inventory to determine whether some of the words are too difficult for the persons he plans to test; he may conclude that the scores on a particular test depend too much on speed for his purposes; or he may notice that an intelligence test developed twenty years ago contains many obsolescent items unsuitable for use today. All these observations about content are relevant to the construct validity of a test. In fact, there is no information provided by any validation procedure that is *not* relevant to construct validity.

The term construct validity was officially introduced into the psychometrist's lexicon in 1954 in the *Technical Recommendations for Psychological Tests and Diagnostic Techniques,* which constituted the first edition of the current APA test *Standards* (1974). Although the validation procedures subsumed under construct validity were not new at the time, the discussions of construct validation that followed served to make the implications of these procedures more explicit and to provide a systematic rationale for their use. Construct validation has focused attention on the role of psychological theory in test construction and on the need to formulate hypotheses that can be proved or disproved in the validation process. It is particularly appropriate in the evaluation of tests for use in research.

In practical contexts, construct validation is suitable for investigating the validity of the criterion measures used in traditional criterion-related test validation (see, e.g., James, 1973). Through an analysis of the correlations of different criterion measures with each other and with other relevant variables, and through factorial analyses of such data, one can learn more about the meaning of a particular criterion. In some instances, the results of such a study may lead to modification or replacement of the criterion chosen to validate a test. Under any circumstances, the results will enrich the interpretation of the test validation study.

Another practical application of construct validation is in the evaluation of tests in situations that do not permit acceptable criterion-related validation studies, as in the local validation of some personnel tests for industrial use. The difficulties encountered in these situations were discussed earlier in this chapter, in connection with synthetic validity. Construct validation offers another alternative approach that could be followed in evaluating the appropriateness of published tests for a particular job. Like synthetic validation, this approach requires a systematic job analysis, followed by a description of worker qualifications expressed in terms of relevant behavioral constructs. If, now, the test has been subjected to sufficient research prior to publication, the data cited in the manual should permit a specification of the principal constructs measured by the test. This information could be used directly in assessing the relevance of the test to the required job functions, if the correspondence of constructs is clear enough; or it could serve as a basis for computing a J-coefficient or some other quantitative index of synthetic validity.

Construct validation has also stimulated the search for novel ways of gathering validity data. Although the principal techniques employed in investigating construct validity have long been familiar, the field of operation has been expanded to admit a wider variety of procedures. This very multiplicity of data-gathering techniques, however, presents certain hazards. It is possible for a test constructor to try a large number of different validation procedures, a few of which will yield positive results by chance. If these confirmatory results were then to be reported without mention of all the validity probes that yielded negative results, a very misleading impression about the validity of a test could be created. Another possible danger in the application of construct validation is that it may open the way for subjective, unverified assertions about test validity. Since construct validity is such a broad and loosely defined concept, it has been widely misunderstood. Some textbook writers and test constructors seem to perceive it as content validity expressed in terms of psychological trait names. Hence, they present as construct validity purely subjective accounts of what they believe (or hope) the test measures.

A further source of possible confusion arises from a statement that construct validation "is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined' " (Cronbach & Meehl, 1955, p. 282). Appearing in the first detailed published analysis of the concept of construct validity, this statement was often incorrectly accepted as justifying a claim for construct validity in the absence of data. That the authors of the statement did not intend such an interpretation is illustrated by their own insistence, in the same article, that "unless the network makes contact with observations . . . construct validation cannot be claimed" (p. 291). In the same connection, they criticize tests for which "a finespun network of rationalizations has been offered as if it were validation" (p. 291). Actually, the theoretical construct, trait, or behavior domain measured by a particular test can be adequately defined only in the light of data gathered in the process of validating that test. Such a definition would take into account the variables with which the test correlated significantly, as well as the conditions found to affect its scores and the groups that differ significantly in such scores. These procedures are entirely in accord with the positive contributions made by the concept of construct validity. It is only through the empirical investigation of the relationships of test scores to other external data that we can discover what a test measures.

CHAPTER 7

Validity: Measurement and Interpretation

CHAPTER 6 was concerned with different concepts of validity and their appropriateness for various testing functions; this chapter deals with quantitative expressions of validity and their interpretation. The test user is concerned with validity at either or both of two stages. First, when considering the suitability of a test for his purposes, he examines available validity data reported in the test manual or other published sources. Through such information, he arrives at a tentative concept of what psychological functions the test actually measures, and he judges the relevance of such functions to his proposed use of the test. In effect, when a test user relies on published validation data, he is dealing with construct validity, regardless of the specific procedures used in gathering the data. As we have seen in Chapter 6, the criteria employed in published studies cannot be assumed to be identical with those the test user wants to predict. Jobs bearing the same title in two different companies are rarely identical. Two courses in freshman English taught in different colleges may be quite dissimilar.

Because of the specificity of each criterion, test users are usually advised to check the validity of any chosen test against local criteria whenever possible. Although published data may strongly suggest that a given test should have high validity in a particular situation, direct corroboration is always desirable. The determination of validity against specific local criteria represents the second stage in the test user's evaluation of validity. The techniques to be discussed in this chapter are especially relevant to the analysis of validity data obtained by the test user himself. Most of them are also useful, however, in understanding and interpreting the validity data reported in test manuals.

VALIDITY COEFFICIENT AND ERROR OF ESTIMATE

MEASUREMENT OF RELATIONSHIP. A validity coefficient is a correlation between test score and criterion measure. Because it provides a single numerical index of test validity, it is commonly used in test manuals to report the validity of a test against each criterion for which data are available. The data used in computing any validity coefficient can also be expressed in the form of an expectancy table or expectancy chart, illustrated in Chapter 4. In fact, such tables and charts provide a convenient way to show what a validity coefficient means for the person tested. It will be recalled that expectancy charts give the probability that an individual who obtains a certain score on the test will attain a specified level of criterion performance. For example, with Table 6 (Ch. 4, p. 101), if we know a student's score on the DAT Verbal Reasoning test, we can look up the chances that he will earn a particular grade in a high school course. The same data yield a validity coefficient of .66 When both test and criterion variables are continuous, as in this example, the familiar Pearson Product-Moment Correlation Coefficient is applicable. Other types of correlation coefficients can be computed when the data are expressed in different forms, as when a two-fold pass-fail criterion is employed (e.g., Fig. 7, Ch. 4). The specific procedures for computing these different kinds of correlations can be found in any standard statistics text.

CONDITIONS AFFECTING VALIDITY COEFFICIENTS. As in the case of reliability, it is essential to specify the *nature of the group* on which a validity coefficient is found. The same test may measure different functions when given to individuals who differ in age, sex, educational level, occupation, or any other relevant characteristic. Persons with different experiential backgrounds, for example, may utilize different work methods to solve the same test problem. Consequently, a test could have high validity in predicting a particular criterion in one population, and little or no validity in another. Or it might be a valid measure of different functions in the two populations. Thus, unless the validation sample is representative of the population on which the test is to be used, validity should be redetermined on a more appropriate sample.

The question of *sample heterogeneity* is relevant to the measurement of validity, as it is to the measurement of reliability, since both characteristics are commonly reported in terms of correlation coefficients. It will be recalled that, other things being equal, the wider the range of scores, the higher will be the correlation. This fact should be kept in mind when interpreting the validity coefficients given in test manuals.

A special difficulty encountered in many validation samples arises from *preselection.* For example, a new test that is being validated for job selection may be administered to a group of newly hired employees on whom criterion measures of job performance will eventually be available. It is likely, however, that such employees represent a superior selection of all those who applied for the job. Hence, the range of such a group in both test scores and criterion measures will be curtailed at the lower end of the distribution. The effect of such preselection will therefore be to lower the validity coefficient. In the subsequent use of the test, when it is administered to all applicants for selection purposes, the validity can be expected to be somewhat higher.

Validity coefficients may also change over time because of changing selection standards. An example is provided by a comparison of validity coefficients computed over a 30-year interval with Yale students (Burnham, 1965). Correlations were found between a predictive index based on College Entrance Examination Board tests and high school records, on the one hand, and average freshman grades, on the other. This correlation dropped from .71 to .52 over the 30 years. An examination of the bivariate distributions clearly reveals the reason for this drop. Because of higher admission standards, the later class was more homogeneous than the earlier class in both predictor and criterion performance. Consequently, the correlation was lower in the later group, although the accuracy with which individuals' grades were predicted showed little change. In other words, the observed drop in correlation did *not* indicate that the predictors were less valid than they had been 30 years earlier. Had the differences in group homogeneity been ignored, it might have been wrongly concluded that this was the case.

For the proper interpretation of a validity coefficient, attention should also be given to the *form of the relationship* between test and criterion. The computation of a Pearson correlation coefficient assumes that the relationship is linear and uniform throughout the range. There is evidence that in certain situations, however, these conditions may not be met (Fisher, 1959; Kahneman & Ghiselli, 1962). Thus, a particular job may require a minimum level of reading comprehension, to enable employees to read instruction manuals, labels, and the like. Once this minimum is exceeded, however, further increments in reading ability may be unrelated to degree of job success. This would be an example of a nonlinear relation between test and job performance. An examination of the bivariate distribution or scatter diagram obtained by plotting reading comprehension scores against criterion measures would show a rise in job performance up to the minimal required reading ability and a leveling off beyond that point. Hence, the entries would cluster around a curve rather than a straight line.

In other situations, the line of best fit may be a straight line, but the individual entries may deviate farther around this line at the upper than at the lower end of the scale. Suppose that performance on a scholastic aptitude test is a necessary but not a sufficient condition for successful achievement in a course. That is, the low-scoring students will perform poorly in the course; but among the high-scoring students, some will perform well in the course and others will perform poorly because of low motivation. In this situation, there will be wider variability of criterion performance among the high-scoring than among the low-scoring students. This condition in a bivariate distribution is known as heteroscedasticity. The Pearson correlation assumes homoscedasticity or equal variability throughout the range of the bivariate distribution. In the present example, the bivariate distribution would be fan-shaped—wide at the upper end and narrow at the lower end. An examination of the bivariate distribution itself will usually give a good indication of the nature of the relationship between test and criterion. Expectancy tables and expectancy charts also correctly reveal the relative effectiveness of the test at different levels.

MAGNITUDE OF A VALIDITY COEFFICIENT. How high should a validity coefficient be? No general answer to this question is possible, since the interpretation of a validity coefficient must take into account a number of concomitant circumstances. The obtained correlation, of course, should be high enough to be *statistically significant* at some acceptable level, such as the .01 or .05 levels discussed in Chapter 5. In other words, before drawing any conclusions about the validity of a test, we should be reasonably certain that the obtained validity coefficient could not have arisen through chance fluctuations of sampling from a true correlation of zero.

Having established a significant correlation between test scores and criterion, however, we need to evaluate the size of the correlation in the light of the uses to be made of the test. If we wish to predict an individual's exact criterion score, such as the grade-point average a student will receive in college, the validity coefficient may be interpreted in terms of the *standard error of estimate,* which is analogous to the error of measurement discussed in connection with reliability. It will be recalled that the error of measurement indicates the margin of error to be expected in an individual's score as a result of the unreliability of the test. Similarly, the error of estimate shows the margin of error to be expected in the individual's predicted criterion score, as a result of the imperfect validity of the test.

The error of estimate is found by the following formula:

$$\sigma_{est.} = \sigma_y\sqrt{1 - r^2_{xy}}$$

in which r^2_{xy} is the square of the validity coefficient and σ_y is the standard deviation of the criterion scores. It will be noted that if the validity were perfect ($r_{xy} = 1.00$), the error of estimate would be zero. On the other hand, with a test having zero validity, the error of estimate is as large as the standard deviation of the criterion distribution ($\sigma_{est.} = \sigma_y\sqrt{1-0} = \sigma_y$). Under these conditions, the prediction is no better than a guess; and the range of prediction error is as wide as the entire distribution of criterion scores. Between these two extremes are to be found the errors of estimate corresponding to tests of varying validity.

Reference to the formula for $\sigma_{est.}$ will show that the term $\sqrt{1 - r^2_{xy}}$ serves to indicate the size of the error *relative to the error that would result from a mere guess*, i.e., with zero validity. In other words, if $\sqrt{1 - r^2_{xy}}$ is equal to 1.00, the error of estimate is as large as it would be if we were to guess the subject's score. The predictive improvement attributable to the use of the test would thus be nil. If the validity coefficient is .80, then $\sqrt{1 - r^2_{xy}}$ is equal to .60, and the error is 60 percent as large as it would be by chance. To put it differently, the use of such a test enables us to predict the individual's criterion performance with a margin of error that is 40 percent smaller than it would be if we were to guess.

It would thus appear that even with a validity of .80, which is unusually high, the error of predicted scores is considerable. If the primary function of psychological tests were to predict each individual's exact position in the criterion distribution, the outlook would be quite discouraging. When examined in the light of the error of estimate, most tests do not appear very efficient. In most testing situations, however, it is not necessary to predict the specific criterion performance of individual cases, but rather to determine which individuals will exceed a certain minimum standard of performance, or cutoff point, in the criterion. What are the chances that Mary Greene will graduate from medical school, that Tom Higgins will pass a course in calculus, or that Beverly Bruce will succeed as an astronaut? Which applicants are likely to be satisfactory clerks, salesmen, or machine operators? Such information is useful not only for group selection but also for individual career planning. For example, it is advantageous for a student to know that he has a good chance of passing all courses in law school, even if we are unable to estimate with certainty whether his grade average will be 74 or 81.

A test may appreciably improve predictive efficiency if it shows *any* significant correlation with the criterion, however low. Under certain circumstances, even validities as low as .20 or .30 may justify inclusion of the test in a selection program. For many testing purposes, evaluation of tests in terms of the error of estimate is unrealistically stringent. Consideration must be given to other ways of evaluating the contribution of a test, which take into account the types of decisions to be made from the scores. Some of these procedures will be illustrated in the following section.

TEST VALIDITY AND DECISION THEORY

BASIC APPROACH. Let us suppose that 100 applicants have been given an aptitude test and followed up until each could be evaluated for success on a certain job. Figure 17 shows the bivariate distribution of test scores and measures of job success for the 100 subjects. The correlation between these two variables is slightly below .70. The minimum acceptable job performance, or criterion cutoff point, is indicated in the diagram by a heavy horizontal line. The 40 cases falling below this line would represent job failures; the 60 above the line, job successes. If all 100 applicants are hired, therefore, 60 percent will succeed on the job. Similarly, if a smaller number were hired at random, without reference to test scores, the proportion of successes would probably be close to 60 percent. Suppose, however, that the test scores are used to select the 45 most promising applicants out of the 100 (selection ratio = .45). In such a case, the 45 individuals falling to the right of the heavy vertical line would be chosen. Within this group of 45, it can be seen that there are 7 job failures, or *false acceptances*, falling below the heavy horizontal line, and 38 job successes. Hence, the percentage of job successes is now 84 rather than 60 (i.e., 38/45 = .84). This increase is attributable to the use of the test as a screening instrument. It will be noted that errors in predicted criterion score that do not affect the decision can be ignored. Only those prediction errors that cross the cutoff line and hence place the individual in the wrong category will reduce the selective effectiveness of the test.

For a complete evaluation of the effectiveness of the test as a screening instrument, another category of cases in Figure 17 must also be examined. This is the category of *false rejections*, comprising the 22 persons who score below the cutoff point on the test but above the criterion cutoff. From these data we would estimate that 22 percent of the total applicant sample are potential job successes who will be lost if the test is used as a screening device with the present cutoff point. These false rejects in a personnel selection situation correspond to the *false positives* in clinical evaluations. The latter term has been adopted from medical practice, in which a test for a pathological condition is reported as positive if the condition is present and negative if the patient is normal. A false positive thus refers to a case in which the test erroneously indicates the presence of a pathological condition, as when brain damage is indicated in an individual who is actually normal. This terminology is likely to be con-

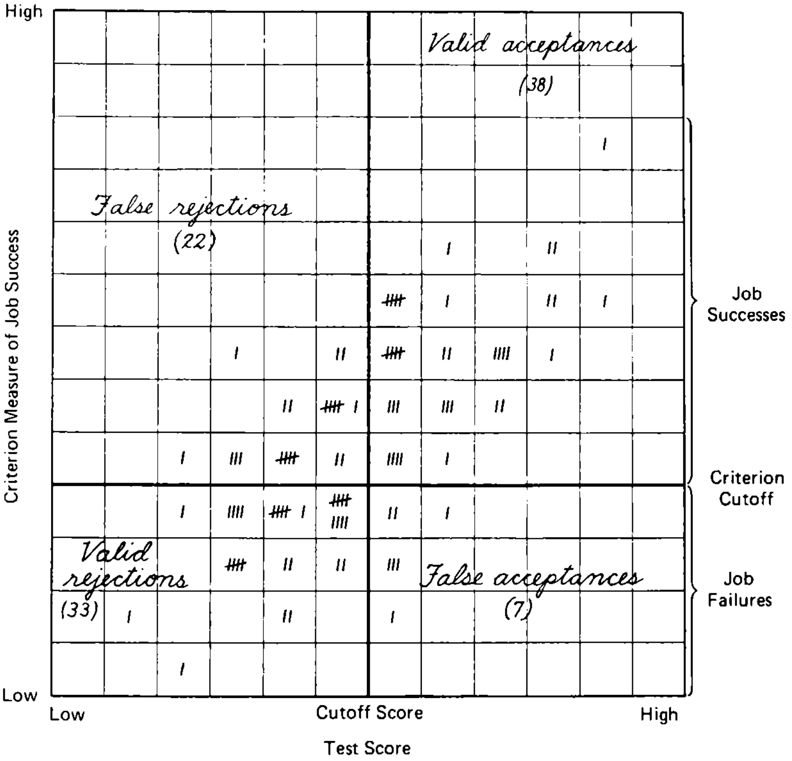


FIG. 17. Increase in the Proportion of "Successes" Resulting from the Use of a Selection Test.

fusing unless we remember that in clinical practice a positive result on a test denotes pathology and unfavorable diagnosis, whereas in personnel selection a positive result conventionally refers to a favorable prediction regarding job performance, academic achievement, and the like.

In setting a cutoff score on a test, attention should be given to the percentage of false rejects (or false positives) as well as to the percentages of successes and failures within the selected group. In certain situations, the cutoff point should be set sufficiently high to exclude all but a few possible failures. This would be the case when the job is of such a nature that a poorly qualified worker could cause serious loss or damage. An example would be a commercial airline pilot. Under other circumstances, it may be more important to admit as many qualified persons as possible, at the risk of including more failures. In the latter case, the number of false rejects can be reduced by the choice of a lower cutoff score. Other factors that normally determine the position of the cutoff score include the available personnel supply, the number of job openings, and the urgency or speed with which the openings must be filled.

In many personnel decisions, the selection ratio is determined by the practical demands of the situation. Because of supply and demand in filling job openings, for example, it may be necessary to hire the top 40 percent of applicants in one case and the top 75 percent in another. When the selection ratio is not externally imposed, the cutting score on a test can be set at that point giving the maximum differentiation between criterion groups. This can be done roughly by comparing the distribution of test scores in the two criterion groups. More precise mathematical procedures for setting optimal cutting scores have also been worked out (Darlington & Stauffer, 1966; Guttman & Raju, 1965; Rorer, Hoffman, La Forge, & Hsieh, 1966). These procedures make it possible to take into account other relevant parameters, such as the relative seriousness of false rejections and false acceptances.

In the terminology of decision theory, the example given in Figure 17 illustrates a simple *strategy*, or plan for deciding which applicants to accept and which to reject. In more general terms, a strategy is a technique for utilizing information in order to reach a decision about individuals. In this case, the strategy was to accept the 45 persons with the highest test scores. The increase in percentage of successful employees from 60 to 84 could be used as a basis for estimating the net benefit resulting from the use of the test.

Statistical decision theory was developed by Wald (1950) with special reference to the decisions required in the inspection and quality control of industrial products. Many of its implications for the construction and interpretation of psychological tests have been systematically worked out by Cronbach and Gleser (1965). Essentially, decision theory is an attempt to put the decision-making process into mathematical form, so that available information may be used to arrive at the most effective decision under specified circumstances. The mathematical procedures employed in decision theory are often quite complex, and few are in a form permitting their immediate application to practical testing problems. Some of the basic concepts of decision theory, however, are proving helpful in the reformulation and clarification of certain questions about tests. A few of these ideas were introduced into testing before the formal development of statistical decision theory and were later recognized as fitting into that framework.

PREDICTION OF OUTCOMES. A precursor of decision theory in psychological testing is to be found in the Taylor-Russell tables (1939), which permit a determination of the net gain in selection accuracy attributable to the use of the test. The information required includes the validity coefficient of the test, the proportion of applicants who must be accepted (selection ratio), and the proportion of successful applicants selected without the use of the test (base rate). A change in any of these three factors can alter the predictive efficiency of the test.

For purposes of illustration, one of the Taylor-Russell tables has been reproduced in Table 14. This table is designed for use when the base rate, or percentage of successful applicants selected prior to the use of the test, is 60. Other tables are provided by Taylor and Russell for other base rates. Across the top of the table are given different values of the selection ratio, and along the side are the test validities. The entries in the body of the table indicate the proportion of successful persons

TABLE 14
Proportion of "Successes" Expected through the Use of Test of Given Validity and Given Selection Ratio, for Base Rate .60.

(From Taylor and Russell, 1939, p. 576)

	Selection Ratio										
Validity	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	.68	.67	.65	.64	.64	.63	.63	.62	.61	.61	.60
.15	.71	.70	.68	.67	.66	.65	.64	.63	.62	.61	.61
.20	.75	.73	.71	.69	.67	.66	.65	.64	.63	.62	.61
.25	.78	.76	.73	.71	.69	.68	.66	.65	.63	.62	.61
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
.35	.85	.82	.78	.75	.73	.71	.69	.67	.65	.63	.62
.40	.88	.85	.81	.78	.75	.73	.70	.68	.66	.63	.62
.45	.90	.87	.83	.80	.77	.74	.72	.69	.66	.64	.62
.50	.93	.90	.86	.82	.79	.76	.73	.70	.67	.64	.62
.55	.95	.92	.88	.84	.81	.78	.75	.71	.68	.64	.62
.60	.96	.94	.90	.87	.83	.80	.76	.73	.69	.65	.63
.65	.98	.96	.92	.89	.85	.82	.78	.74	.70	.65	.63
.70	.99	.97	.94	.91	.87	.84	.80	.75	.71	.66	.63
.75	.99	.99	.96	.93	.90	.86	.81	.77	.71	.66	.63
.80	1.00	.99	.98	.95	.92	.88	.83	.78	.72	.66	.63
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

selected after the use of the test. Thus, the difference between .60 and any one table entry shows the increase in proportion of successful selections attributable to the test.

Obviously if the selection ratio were 100 percent, that is, if all applicants had to be accepted, no test, however valid, could improve the selection process. Reference to Table 14 shows that, when as many as 95 percent of applicants must be admitted, even a test with perfect validity ($r = 1.00$) would raise the proportion of successful persons by only 3 percent (.60 to .63). On the other hand, when only 5 percent of applicants need to be chosen, a test with a validity coefficient of only .30 can raise the percentage of successful applicants selected from 60 to 82. The rise from 60 to 82 represents the *incremental validity* of the test (Sechrest, 1963), or the increase in predictive validity attributable to the test. It indicates the contribution the test makes to the selection of individuals who will meet the minimum standards in criterion performance. In applying the Taylor-Russell tables, of course, test validity should be computed on the same sort of group used to estimate percentage of prior successes. In other words, the contribution of the test is not evaluated against chance success unless applicants were previously selected by chance—a most unlikely circumstance. If applicants had been selected on the basis of previous job history, letters of recommendation, and interviews, the contribution of the test should be evaluated on the basis of what the test adds to these previous selection procedures.

The incremental validity resulting from the use of a test depends not only on the selection ratio but also on the base rate. In the previously illustrated job selection situation, the base rate refers to the proportion of successful employees prior to the introduction of the test for selection purposes. Table 14 shows the anticipated outcomes when the base rate is .60. For other base rates, we need to consult the other appropriate tables in the cited reference (Taylor & Russell, 1939). Let us consider an example in which test validity is .40 and the selection ratio is 70 percent. Under these conditions, what would be the contribution or incremental validity of the test if we begin with a base rate of 50 percent? And what would be the contribution if we begin with more extreme base rates of 10 and 90 percent? Reference to the appropriate Taylor-Russell tables for these base rates shows that the percentage of successful employees would rise from 50 to 75 in the first case; from 10 to 21 in the second; and from 90 to 99 in the third. Thus, the improvement in percentage of successful employees attributable to the use of the test is 25 when the base rate was 50, but only 11 and 9 when the base rates were more extreme.

The implications of extreme base rates are of special interest in clinical psychology, where the base rate refers to the frequency of the pathological condition to be diagnosed in the population tested (Buchwald, 1965; Cureton, 1957a; Meehl & Rosen, 1955; J. S. Wiggins, 1973). For example, if 5 percent of the intake population of a clinic has organic brain damage, then 5 percent is the base rate of brain damage in this population. Although the introduction of any valid test will improve predictive or diagnostic accuracy, the improvement is greatest when the base rates are closest to 50 percent. With the extreme base rates found with rare pathological conditions, however, the improvement may be negligible. Under these conditions, the use of a test may prove to be unjustified when the cost of its administration and scoring is taken into account. In a clinical situation, this cost would include the time of professional personnel that might otherwise be spent on the treatment of additional cases (Buchwald, 1965). The number of false positives, or normal individuals incorrectly classified as pathological, would of course increase this overall cost in a clinical situation.

When the seriousness of a rare condition makes its diagnosis urgent, tests of moderate validity may be employed in an early stage of sequential decisions. For example, all cases might first be screened with an easily administered test of moderate validity. If the cutoff score is set high enough (high scores being favorable), there will be few false negatives but many false positives, or normals diagnosed as pathological. The latter can then be detected through a more intensive individual examination given to all cases diagnosed as positive by the test. This solution would be appropriate, for instance, when available facilities make the intensive individual examination of all cases impracticable.

RELATION OF VALIDITY TO MEAN OUTPUT LEVEL. In many practical situations, what is wanted is an estimate of the effect of the selection test, not on percentage of persons exceeding the minimum performance, but on overall *output* of the selected persons. How does the actual level of job proficiency or criterion achievement of the workers hired on the basis of the test compare with that of the total applicant sample that would have been hired without the test? Following the work of Taylor and Russell, several investigators addressed themselves to this question (Brogden, 1946; Brown & Ghiselli, 1953; Jarrett, 1948; Richardson, 1944). Brogden (1946) first demonstrated that the expected increase in output is directly proportional to the validity of the test. Thus, the improvement resulting from the use of a test of validity .50 is 50 percent as great as the improvement expected from a test of perfect validity.

The relation between test validity and expected rise in criterion achievement can be readily seen in Table 15.[1] Expressing criterion scores

---

[1] A table including more values for both selection ratios and validity coefficients was prepared by Naylor and Shine (1965).

TABLE 15

Mean Standard Criterion Score of Accepted Cases in Relation to Test Validity and Selection Ratio

(From Brown and Ghiselli, 1953, p. 342)

Selection Ratio	Validity Coefficient																				
	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
.05	.00	.10	.21	.31	.42	.52	.62	.73	.83	.94	1.04	1.14	1.25	1.35	1.46	1.56	1.66	1.77	1.87	1.98	2.08
.10	.00	.09	.18	.26	.35	.44	.53	.62	.70	.79	.88	.97	1.05	1.14	1.23	1.32	1.41	1.49	1.58	1.67	1.76
.15	.00	.08	.15	.23	.31	.39	.46	.54	.62	.70	.77	.85	.93	1.01	1.08	1.16	1.24	1.32	1.39	1.47	1.55
.20	.00	.07	.14	.21	.28	.35	.42	.49	.56	.63	.70	.77	.84	.91	.98	1.05	1.12	1.19	1.26	1.33	1.40
.25	.00	.06	.13	.19	.25	.32	.38	.44	.51	.57	.63	.70	.76	.82	.89	.95	1.01	1.08	1.14	1.20	1.27
.30	.00	.06	.12	.17	.23	.29	.35	.40	.46	.52	.58	.64	.69	.75	.81	.87	.92	.98	1.04	1.10	1.16
.35	.00	.05	.11	.16	.21	.26	.32	.37	.42	.48	.53	.58	.63	.69	.74	.79	.84	.90	.95	1.00	1.06
.40	.00	.05	.10	.15	.19	.24	.29	.34	.39	.44	.48	.53	.58	.63	.68	.73	.77	.82	.87	.92	.97
.45	.00	.04	.09	.13	.18	.22	.26	.31	.35	.40	.44	.48	.53	.57	.62	.66	.70	.75	.79	.84	.88
.50	.00	.04	.08	.12	.16	.20	.24	.28	.32	.36	.40	.44	.48	.52	.56	.60	.64	.68	.72	.76	.80
.55	.00	.04	.07	.11	.14	.18	.22	.25	.29	.32	.36	.40	.43	.47	.50	.54	.58	.61	.65	.68	.72
.60	.00	.03	.06	.10	.13	.16	.19	.23	.26	.29	.32	.35	.39	.42	.45	.48	.52	.55	.58	.61	.64
.65	.00	.03	.06	.09	.11	.14	.17	.20	.23	.26	.28	.31	.34	.37	.40	.43	.46	.48	.51	.54	.57
.70	.00	.02	.05	.07	.10	.12	.15	.17	.20	.22	.25	.27	.30	.32	.35	.37	.40	.42	.45	.47	.50
.75	.00	.02	.04	.06	.08	.11	.13	.15	.17	.19	.21	.23	.25	.27	.30	.32	.33	.36	.38	.40	.42
.80	.00	.02	.04	.05	.07	.09	.11	.12	.14	.16	.18	.19	.21	.22	.25	.26	.28	.30	.32	.33	.35
.85	.00	.01	.03	.04	.05	.07	.08	.10	.11	.12	.14	.15	.16	.18	.19	.20	.22	.23	.25	.26	.27
.90	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20
.95	.00	.01	.01	.02	.02	.03	.03	.04	.04	.05	.05	.06	.07	.07	.08	.08	.09	.09	.10	.10	.11

as standard scores with a mean of zero and an *SD* of 1.00, this table gives the expected mean criterion score of workers selected with a test of given validity and with a given selection ratio. In this context, the base output mean, corresponding to the performance of applicants selected without use of the test, is given in the column for zero validity. Using a test with zero validity is equivalent to using no test at all. To illustrate the use of the table, let us assume that the highest scoring 20 percent of the applicants are hired (selection ratio = .20) by means of a test whose validity coefficient is .50. Reference to Table 15 shows that the mean criterion performance of this group is .70 *SD* above the expected base mean of an untested sample. With the same 20 percent selection ratio and a perfect test (validity coefficient = 1.00), the mean criterion score of the accepted applicants would be 1.40, just twice what it would be with the test of validity .50. Similar direct linear relations will be found if other mean criterion performances are compared within any row of Table 15. For instance, with a selection ratio of 60 percent, a validity of .25 yields a mean criterion score of .16, while a validity of .50 yields a mean of .32. Again, doubling the validity doubles the output rise.

The evaluation of test validity in terms of either mean predicted output or proportion of persons exceeding a minimum criterion cutoff is obviously much more favorable than an evaluation based on the previously discussed error of estimate. The reason for the difference is that prediction errors that do not affect decisions are irrelevant to the selection situation. For example, if Smith and Jones are both superior workers and are both hired on the basis of the test, it does not matter if the test shows Smith to be better than Jones while in job performance Jones excels Smith.

THE ROLE OF VALUES IN DECISION THEORY. It is characteristic of decision theory that tests are evaluated in terms of their effectiveness in a specific situation. Such evaluation takes into account not only the validity of the test in predicting a particular criterion but also a number of other parameters, including base rate and selection ratio. Another important parameter is the relative *utility* of expected outcomes, the judged favorableness or unfavorableness of each outcome. The lack of adequate systems for assigning values to outcomes in terms of a uniform utility scale is one of the chief obstacles to the application of decision theory. In industrial decisions, a dollar-and-cents value can frequently be assigned to different outcomes. Even in such cases, however, certain outcomes pertaining to good will, public relations, and employee morale are difficult to assess in monetary terms. Educational decisions must take into account institutional goals, social values, and other relatively intangible factors. Individual decisions, as in counseling, must consider the individual's preferences and value system. It has been repeatedly pointed out, however, that decision theory did not introduce the problem of values into the decision process, but merely made it explicit. Value systems have always entered into decisions, but they were not heretofore clearly recognized or systematically handled.

In choosing a decision strategy, the goal is to maximize expected utilities across all outcomes. Reference to the schematic representation of a simple decision strategy in Figure 18 will help to clarify the procedure. This diagram shows the decision strategy illustrated in Figure 17, in which a single test is administered to a group of applicants and the decision to accept or reject an applicant is made on the basis of a cutoff score on the test. There are four possible outcomes, including valid and false acceptances and valid and false rejections. The probability of each outcome can be found from the number of persons in each of the four sections of Figure 17. Since there were 100 applicants in that example, these numbers divided by 100 give the probabilities of the four outcomes listed in Figure 18. The other data needed are the utilities of the different outcomes, expressed on a common scale. The expected overall utility of the strategy could then be found by multiplying the probability of each outcome by the utility of the outcome, adding these products for the four outcomes, and subtracting a value corresponding to the cost of testing.[2] This last term highlights the fact that a test of low validity is more likely to be retained if it is short, inexpensive, easily administered by relatively untrained personnel, and suitable for group administration. An individual test requiring a trained examiner or expensive equipment would need a higher validity to justify its use.

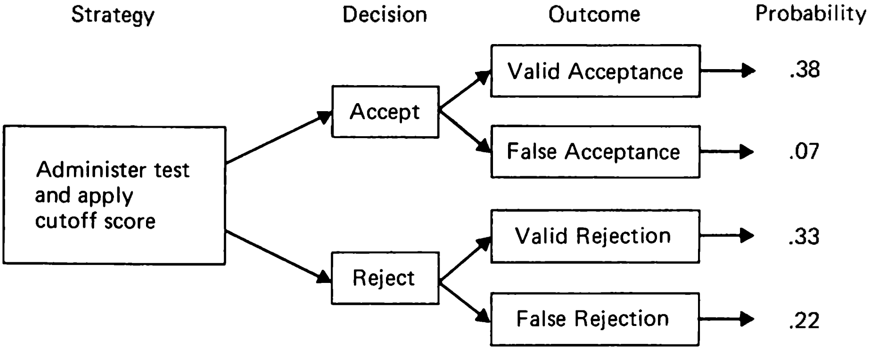


FIG. 18. A Simple Decision Strategy.

---

[2] For a fictitious example illustrating all steps in these computations, see J. S. Wiggins (1973), pp. 257–274.

SEQUENTIAL STRATEGIES AND ADAPTIVE TREATMENTS. In some situations, the effectiveness of a test may be increased through the use of more complex decision strategies which take still more parameters into account. Two examples will serve to illustrate these possibilities. First, tests may be used to make *sequential* rather than terminal decisions. With the simple decision strategy illustrated in Figures 17 and 18, all decisions to accept or reject are treated as terminal. Figure 19, on the other hand, shows a two-stage sequential decision. Test A could be a short and easily administered screening test. On the basis of performance on this test, individuals would be sorted into three categories, including those clearly accepted or rejected, as well as an intermediate "uncertain" group to be examined further with more intensive techniques, represented by Test B. On the basis of the second-stage testing, this group would be sorted into accepted and rejected categories.

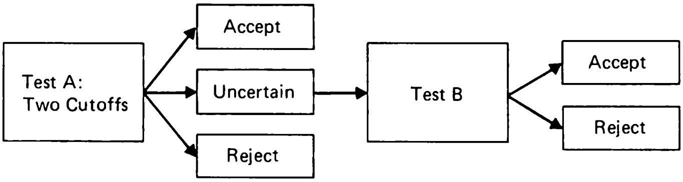


FIG. 19. A Sequential Decision Strategy.

Such sequential testing can also be employed within a single testing session, to maximize the effective use of testing time (DeWitt & Weiss, 1974; Linn, Rock, & Cleary, 1969; Weiss & Betz, 1973). Although applicable to paper-and-pencil printed group tests, sequential testing is particularly well suited for computer testing. Essentially the sequence of items or item groups within the test is determined by the examinee's own performance. For example, everyone might begin with a set of items of intermediate difficulty. Those who score poorly are routed to easier items; those who score well, to more difficult items. Such branching may occur repeatedly at several stages. The principal effect is that each examinee attempts only those items suited to his ability level, rather than trying all items. Sequential testing models will be discussed further in Chapter 11, in connection with the utilization of computers in group testing.

Another strategy, suitable for the diagnosis of psychological disorders, is to use only two categories, but to test further *all* cases classified as positives (i.e., possibly pathological) by the preliminary screening test. This is the strategy cited earlier in this section, in connection with the use of tests to diagnose pathological conditions with very low base rates.

It should also be noted that many personnel decisions are in effect sequential, although they may not be so perceived. Incompetent employees hired because of prediction errors can usually be discharged after a probationary period; failing students can be dropped from college at several stages. In such situations, it is only adverse selection decisions that are terminal. To be sure, incorrect selection decisions that are later rectified may be costly in terms of several value systems. But they are often less costly than terminal wrong decisions.

A second condition that may alter the effectiveness of a psychological test is the availability of alternative treatments and the possibility of *adapting treatments* to individual characteristics. An example would be the utilization of different training procedures for workers at different aptitude levels, or the introduction of compensatory educational programs for students with certain educational disabilities. Under these conditions, the decision strategy followed in individual cases should take into account available data on the interaction of initial test score and differential treatment. When adaptive treatments are utilized, the success rate is likely to be substantially improved. Because the assignment of individuals to alternative treatments is essentially a classification rather than a selection problem, more will be said about the required methodology in a later section on classification decisions.

The examples cited illustrate a few of the ways in which the concepts and rationale of decision theory can assist in the evaluation of psychological tests for specific testing purposes. Essentially, decision theory has served to focus attention on the complexity of factors that determine the contribution a given test can make in a particular situation. The validity coefficient alone cannot indicate whether or not a test should be used, since it is only one of the factors to be considered in evaluating the impact of the test on the efficacy of the total decision process.[3]

MODERATOR VARIABLES

DIFFERENTIALLY PREDICTABLE SUBSETS OF PERSONS. The validity of a test for a given criterion may vary among subgroups differing in personal characteristics. The classic psychometric model assumes that prediction errors are characteristic of the test rather than of the person and that these errors are randomly distributed among persons. With the flexibility of approach ushered in by decision theory, there has been increasing exploration of prediction models involving interaction between persons and tests. Such interaction implies that the same test may be a better predictor for certain classes or subsets of persons than it is for others. For example, a given test may be a better predictor of criterion performance for men than for women, or a better predictor for applicants from a lower than for applicants from a higher socioeconomic level. In these examples, sex and socioeconomic level are known as *moderator variables,* since they moderate the validity of the test (Saunders, 1956).

---

[3] For a fuller discussion of the implications of decision theory for test use, see J. S. Wiggins (1973), Ch. 6, and at a more technical level, Cronbach and Gleser (1965).

When computed in a total group, the validity coefficient of a test may be too low to be of much practical value in prediction. But when recomputed in subsets of individuals differing in some identifiable characteristic, validity may be high in one subset and negligible in another. The test could thus be used effectively in making decisions regarding persons in the first group but not in the second. Perhaps another test or some other assessment device could be found that is an effective predictor in the second group.

A moderator variable is some characteristic of persons that makes it possible to predict the predictability of different individuals with a given instrument. It may be a demographic variable, such as sex, age, educational level, or socioeconomic background; or it may be a score on another test. Interests and motivation often function as moderator variables. Thus, if an applicant has little interest in a job, he will probably perform poorly regardless of his scores on relevant aptitude tests. Among such persons, the correlation between aptitude test scores and job performance would be low. For individuals who are interested and highly motivated, on the other hand, the correlation between aptitude test score and job success may be quite high.

EMPIRICAL EXAMPLES OF MODERATOR VARIABLES. Evidence for the operation of moderator variables comes from a variety of sources. In a survey of several hundred correlation coefficients between aptitude test scores and academic grades, H. G. Seashore (1962) found higher correlations for women than for men in the large majority of instances. The same trend was found in high school and college, although the trend was more pronounced at the college level. The data do not indicate the reason for this sex difference in the predictability of academic achievement, but it may be interesting to speculate about it in the light of other known sex differences. If women students in general tend to be more conforming and more inclined to accept the values and standards of the school situation, their class achievement will probably depend largely on their abilities. If, on the other hand, men students tend to concentrate their efforts on those activities (in or out of school) that arouse their individual interests, these interest differences would introduce additional variance in their course achievement and would make it more difficult to predict achievement from test scores. Whatever the reason for the difference, sex does appear to function as a moderator variable in the predictability of academic grades from aptitude test scores.

A number of investigations have been specially designed to assess the role of moderator variables in the prediction of academic achievement. Several studies (Frederiksen & Gilbert, 1960; Frederiksen & Melville, 1954; Stricker, 1966) tested the hypothesis that the more compulsive students, identified through two tests of compulsivity, would put a great deal of effort into their course work, regardless of their interest in the courses, but that the effort of the less compulsive students would depend on their interest. Since effort will be reflected in grades, the correlation between the appropriate interest test scores and grades should be higher among noncompulsive than among compulsive students. This hypothesis was confirmed in several groups of male engineering students, but not among liberal arts students of either sex. Moreover, lack of agreement among different indicators of compulsivity casts doubt on the generality of the construct that was being measured.

In another study (Grooms & Endler, 1960), the college grades of the more anxious students correlated higher ($r = .63$) with aptitude and achievement test scores than did the grades of the less anxious students ($r = .19$). A different approach is illustrated by Berdie (1961), who investigated the relation between intraindividual variability on a test and the predictive validity of the same test. It was hypothesized that a given test will be a better predictor for those individuals who perform more consistently in different parts of the test—and whose total scores are thus more reliable. Although the hypothesis was partially confirmed, the relation proved to be more complex than anticipated (Berdie, 1969).

In a different context, there is evidence that self-report personality inventories may have higher validity for some types of neurotics than for others (Fulkerson, 1959). The characteristic behavior of the two types tends to make one type careful and accurate in reporting symptoms, the other careless and evasive. The individual who is characteristically precise and careful about details, who tends to worry about his problems, and who uses intellectualization as a primary defense is likely to provide a more accurate picture of his emotional difficulties on a self-report inventory than is the impulsive, careless individual who tends to avoid expressing unpleasant thoughts and emotions and who uses denial as a primary defense.

Ghiselli (1956, 1960a, 1960b, 1963, 1968; Ghiselli & Sanders, 1967) has extensively explored the role of moderator variables in industrial situations. In a study of taxi drivers (Ghiselli, 1956), the correlation between an aptitude test and a job-performance criterion in the total applicant sample was only .220. The group was then sorted into thirds on the basis of scores on an occupational interest test. When the validity of the aptitude test was recomputed within the third whose occupational interest level was most appropriate for the job, it rose to .664.

A technique employed by Ghiselli in much of his research consists in finding for each individual the absolute difference (*D*) between his actual and his predicted criterion scores. The smaller the value of *D*, the more predictable is the individual's criterion score. A predictability scale is then developed by comparing the item responses of two contrasted subgroups selected on the basis of their *D* scores. The predictability scale is subsequently applied to a new sample, to identify highly predictable and poorly predictable subgroups, and the validity of the original test is compared in these two subgroups. This approach has shown considerable promise as a means of identifying persons for whom a test will be a good or a poor predictor. An extension of the same procedure has been developed to determine in advance which of two tests will be a better predictor for each individual (Ghiselli, 1960a).

Other investigators (Dunnette, 1972; Hobert & Dunnette, 1967) have argued that Ghiselli's *D* index, based on the absolute amount of prediction error without regard to direction of error, may obscure important individual differences. Alternative procedures, involving separate analyses of overpredicted and underpredicted cases, have accordingly been proposed.

At this time, the identification and use of moderator variables are still in an exploratory phase. Considerable caution is required to avoid methodological pitfalls (see, e.g., Abrahams & Alf, 1972a, 1972b; Dunnette, 1972; Ghiselli, 1972; Velicer, 1972a, 1972b). The results are usually quite specific to the situations in which they were obtained. And it is important to check the extent to which the use of moderators actually improves the prediction that could be achieved through other more direct means (Pinder, 1973).

COMBINING INFORMATION FROM DIFFERENT TESTS

For the prediction of practical criteria, not one but several tests are generally required. Most criteria are complex, the criterion measure depending on a number of different traits. A single test designed to measure such a criterion would thus have to be highly heterogeneous. It has already been pointed out, however, that a relatively homogeneous test, measuring largely a single trait, is more satisfactory because it yields less ambiguous scores (Ch. 5). Hence, it is usually preferable to use a combination of several relatively homogeneous tests, each covering a different aspect of the criterion, rather than a single test consisting of a hodgepodge of many different kinds of items.

When a number of specially selected tests are employed together to predict a single criterion, they are known as a *test battery*. The chief problem arising in the use of such batteries concerns the way in which scores on the different tests are to be combined in arriving at a decision regarding each individual. The statistical procedures followed for this purpose are of two major types, namely, multiple regression equation and multiple cutoff scores.

When tests are administered in the intensive study of individual cases, as in clinical diagnosis, counseling, or the evaluation of high-level executives, it is a common practice for the examiner to utilize test scores without further statistical analysis. In preparing a case report and in making recommendations, the examiner relies on judgment, past experience, and theoretical rationale to interpret score patterns and integrate findings from different tests. Such clinical use of test scores will be discussed further in Chapter 16.

MULTIPLE REGRESSION EQUATION. The multiple regression equation yields a predicted criterion score for each individual on the basis of his scores on all the tests in the battery. The following regression equation illustrates the application of this technique to predicting a student's achievement in high school mathematics courses from his scores on verbal (V), numerical (N), and reasoning (R) tests:

$$\text{Mathematics Achievement} = .21\,V + .21\,N + .32\,R + 1.35$$

In this equation, the student's stanine score on each of the three tests is multiplied by the corresponding weight given in the equation. The sum of these products, plus a constant (1.35), gives the student's predicted stanine position in mathematics courses.

Suppose that Bill Jones receives the following stanine scores:

Verbal	6
Numerical	4
Reasoning	8

The estimated mathematics achievement of this student is found as follows:

$$\text{Math. Achiev.} = (.21)(6) + (.21)(4) + (.32)(8) + 1.35 = 6.01$$

Bill's predicted stanine is approximately 6. It will be recalled (Ch. 4) that a stanine of 5 represents average performance. Bill would thus be expected to do somewhat better than average in mathematics courses. His very superior performance in the reasoning test ($R = 8$) and his above-average score on the verbal test ($V = 6$) compensate for his poor score in speed and accuracy of computation ($N = 4$).

Specific techniques for the computation of regression equations can be found in texts on psychological statistics (e.g., Guilford and Fruchter, 1973). Essentially, such an equation is based on the correlation of each test with the criterion, as well as on the intercorrelations among the tests. Obviously, those tests that correlate higher with the criterion should receive more weight. It is equally important, however, to take into account the correlation of each test with the other tests in the battery. Tests correlating highly with each other represent needless duplication, since they cover to a large extent the same aspects of the criterion. The inclusion of two such tests will not appreciably increase the validity of the entire battery, even though both tests may correlate highly with the criterion. In such a case, one of the tests would serve about as effectively as the pair; only one would therefore be retained in the battery.

Even after the most serious instances of duplication have been eliminated, however, the tests remaining in the battery will correlate with each other to varying degrees. For maximum predictive value, tests that make a more nearly unique contribution to the total battery should receive greater weight than those that partly duplicate the functions of other tests. In the computation of a multiple regression equation, each test is weighted in direct proportion to its correlation with the criterion and in inverse proportion to its correlations with the other tests. Thus, the highest weight will be assigned to the test with the highest validity and the least amount of overlap with the rest of the battery.

The validity of the entire battery can be found by computing the multiple correlation (R) between the criterion and the battery. This correlation indicates the highest predictive value that can be obtained from the given battery, when each test is given optimum weight for predicting the criterion in question. The optimum weights are those determined by the regression equation.

It should be noted that these weights are optimum only for the particular sample in which they were found. Because of chance errors in the correlation coefficients used in deriving them, the regression weights may vary from sample to sample. Hence, the battery should be cross-validated by correlating the predicted criterion scores with the actual criterion scores in a new sample. Formulas are available for estimating the amount of shrinkage in a multiple correlation to be expected when the regression equation is applied to a second sample, but empirical verification is preferable whenever possible. The larger the sample on which regression weights were derived, the smaller the shrinkage will be.

Under certain conditions the predictive validity of a battery can be improved by including in the regression equation a test having a zero correlation with the criterion but a high correlation with another test in the battery. This curious situation arises when the test that is uncorrelated with the criterion acts as a *suppressor variable* to eliminate or suppress the irrelevant variance in the other test. For example, reading comprehension might correlate highly with scores on a mathematical or a mechanical aptitude test, because the test problems require the understanding of complicated written instructions. If reading comprehension is irrelevant to the job behavior to be predicted, the reading comprehension required by the tests introduces error variance and lowers the predictive validity of the tests. Administering a reading comprehension test and including scores on this test in the regression equation will eliminate this error variance and raise the validity of the battery. The suppressor variable appears in the regression equation with a negative weight. Thus, the higher an individual's score on reading comprehension, the more is deducted from his score on the mathematical or mechanical test.

The use of suppressor variables is illustrated by a study of 63 industrial mechanics (Sorenson, 1966). The most effective battery for predicting job performance in this group included: (1) a questionnaire covering education, previous mechanical experience, and other background data (criterion correlation = .30); (2) a mechanical insight test stressing practical mechanics of the "nuts-and-bolts" type (criterion correlation = .22); and (3) a test of mechanical comprehension oriented toward the academic understanding of mechanical principles (criterion correlation = −.04; correlation with test 2 = .71). The third test functioned as a suppressor variable, as can be seen from the following regression equation:

$$C = 17T_1 + 10T_2 - 6T_3 + 866$$

Without the suppressor variable the test battery would "overpredict" the job performance of individuals who obtain high scores on the practical mechanics test through an application of mechanical principles but who lack the practical mechanical know-how required on the job. The irrelevant contribution of academic knowledge of mechanical principles to scores on the practical mechanics tests was thus ruled out by the suppressor variable.

Attempts to employ suppressor variables to improve the validity of personality tests have proved disappointing (J. S. Wiggins, 1973). In any situation, moreover, the more direct procedure of revising a test to eliminate the irrelevant variance is preferable to the indirect statistical elimination of such variance through a suppressor variable. When changes in the test are not feasible, the investigation of suppressor variables should be considered.[4]

[4] There has also been some exploration of the inclusion of continuous moderator variables in regression equations through nonadditive and higher-order functions; but the results have not been promising (Kirkpatrick et al., 1968; Saunders, 1956; J. S. Wiggins, 1973).

MULTIPLE CUTOFF SCORES. An alternative strategy for combining test scores utilizes multiple cutoff points. Briefly, this procedure involves the establishment of a minimum cutoff score on each test. Every individual who falls below such a minimum score on *any one* of the tests is rejected. Only those persons who reach or exceed the cutoff scores in all tests are accepted. An example of this technique is provided by the General Aptitude Test Battery (GATB) developed by the United States Employment Service for use in the occupational counseling program of its State Employment Service offices (U.S. Department of Labor, 1970a). Of the nine aptitude scores yielded by this battery, those to be considered for each occupation were chosen on the basis of criterion correlations as well as means and standard deviations of workers in that occupation.

TABLE 16
Illustrative Data Used to Establish Cutoff Scores on GATB

(From U.S. Department of Labor, 1958, p. 10)

	Aptitude	Mean	*SD*	Criterion Correlation
G	General Learning Ability	75.1	14.2	−.094
V	Verbal	80.1	11.3	−.085
N	Numerical	73.2	18.4	−.064
S	Spatial	78.9	15.9	.041
P	Form Perception	80.1	23.5	−.012
Q	Clerical Perception	86.3	16.6	.088
K	Motor Coordination	89.3	20.7	.316*
F	Finger Dexterity	92.4	18.1	.155
M	Manual Dexterity	88.2	18.6	.437**

* Significant at .05 level.
** Significant at .01 level.

The development of GATB occupational standards for machine cutters in the food-canning and preserving industry is illustrated in Table 16. In terms of standard scores with a mean of 100 and an *SD* of 20, the cutoff scores for this occupation were set at 75 in Motor Coordination (*K*), Finger Dexterity (*F*), and Manual Dexterity (*M*) (U.S. Department of Labor, 1970a, Section IV, p. 51). Table 16 gives mean, standard deviation, and correlation with the criterion (supervisory ratings) for each of the nine scores in a group of 57 women workers. On the basis of criterion correlations, Manual Dexterity and Motor Coordination appeared promising. Finger Dexterity was added because it yielded the highest mean score in the battery, even though individual differences within the group were not significantly correlated with criterion ratings. It would seem that women who enter or remain in this type of job are already selected with regard to Finger Dexterity.[5]

The validity of the composite *KFM* pattern of cutting scores in a group of 194 workers is shown in Table 17. It will be seen that, of 150 good workers, 120 fell above the cutting scores in the three aptitudes and 30 were false rejects, falling below one or more cutoffs. Of the 44 poor workers, 30 were correctly identified and 14 were false acceptances. The overall efficacy of this cutoff pattern is indicated by a tetrachoric correlation of .70 between predicted status and criterion ratings.

TABLE 17
Effectiveness of GATB Cutoff Scores on Aptitudes *K*, *F*, and *M* in Identifying Good and Poor Workers

(From U.S. Department of Labor, 1958, p. 14)

Criterion Rating	Score Pattern		Total
	Nonqualifying	Qualifying	
Good	30	120	150
Poor	30	14	44
Total	60	134	194

If only scores yielding significant validity coefficients are taken into account, one or more essential abilities in which all workers in the occupation excel might be overlooked. Hence the need for considering also those aptitudes in which workers excel as a group, even when individual differences beyond a certain minimum are unrelated to degree of job success. The multiple cutoff method is preferable to the regression equation in situations such as these, in which test scores are not linearly related to the criterion. In some jobs, moreover, workers may be so homogeneous in a key trait that the range of individual differences is too narrow to yield a significant correlation between test scores and criterion.

The strongest argument for the use of multiple cutoffs rather than a regression equation centers around the question of compensatory qualifications. With the regression equation, an individual who rates low in one test may receive an acceptable total score because he rates very high in some other test in the battery. A marked deficiency in one skill may thus be compensated for by outstanding ability along other lines. It is possible, however, that certain types of activity may require essential skills for which there is no substitute. In such cases, individuals falling below the required minimum in the essential skill will fail, regardless of their other abilities. An opera singer, for example, cannot afford to have poor pitch discrimination, regardless of how well he meets the other requirements of such a career. Similarly, operators of sound-detection devices in submarines need good auditory discrimination. Those men incapable of making the necessary discriminations cannot succeed in such an assignment, regardless of superior mechanical aptitude, general intelligence, or other traits in which they may excel. With a multiple cutoff strategy, individuals lacking any essential skill would always be rejected, while with a regression equation they might be accepted.

---

[5] The data have been somewhat simplified for illustrative purposes. Actually, the final choice of aptitudes and cutting scores was based on separate analyses of three groups of workers on related jobs, on the results obtained in a combined sample of 194 cases, and on qualitative job analyses of the operations involved.

When the relation between tests and criterion is linear and additive, on the other hand, a higher proportion of correct decisions will be reached with a regression equation than with multiple cutoffs. Another important advantage of the regression equation is that it provides an estimate of each person's criterion score, thereby permitting the relative evaluation of all individuals. With multiple cutoffs, no further differentiation is possible among those accepted or among those rejected. In many situations, the best strategy may involve a combination of both procedures. Thus, the multiple cutoff may be applied first, in order to reject those falling below minimum standards on any test, and predicted criterion scores may then be computed for the remaining acceptable cases by the use of a regression equation. If enough is known about the particular job requirements, the preliminary screening may be done in terms of only one or two essential skills, prior to the application of the regression equation.

USE OF TESTS FOR CLASSIFICATION DECISIONS

THE NATURE OF CLASSIFICATION. Psychological tests may be used for purposes of selection, placement, or classification. In *selection*, each individual is either accepted or rejected. Deciding whether or not to admit a student to college, to hire a job applicant, or to accept an army recruit for officer training are examples of selection decisions. When selection is done sequentially, the earlier stages are often called "screening," the term "selection" being reserved for the more intensive final stages. "Screening" may also be used to designate any rapid, rough selection process even when not followed by further selection procedures.

Both placement and classification differ from selection in that no one is rejected, or eliminated from the program. All individuals are assigned to appropriate "treatments" so as to maximize the effectiveness of outcomes.

In *placement,* the assignments are based on a single score. This score may be derived from a single test, such as a mathematics achievement test. If a battery of tests has been administered, a composite score computed from a single regression equation would be employed. Examples of placement decisions include the sectioning of college freshmen into different mathematics classes on the basis of their achievement test scores, assigning applicants to clerical jobs requiring different levels of skill and responsibility, and placing psychiatric patients into "more disturbed" and "less disturbed" wards. It is evident that in each of these decisions only one criterion is employed and that placement is determined by the individual's position along a single predictor scale.

Classification, on the other hand, always involves two or more criteria. In a military situation, for example, classification is a major problem, since each man in an available manpower pool must be assigned to the military specialty where he can serve most effectively. Classification decisions are likewise required in industry, when new employees are assigned to training programs for different kinds of jobs. Other examples include the counseling of students regarding choice of college curriculum (science, liberal arts, etc.), as well as field of concentration. Counseling is based essentially on classification, since the client is told his chances of succeeding in different academic programs or occupations. Clinical diagnosis is likewise a classification problem, the major purposes of each diagnosis being a decision regarding the most appropriate type of therapy.

Although placement can be done with either one or more predictors, classification requires multiple predictors whose validity is individually determined against each criterion. A classification battery requires a different regression equation for each criterion. Some of the tests may have weights in all the equations, although of different values; others may be included in only one or two equations, having zero or negligible weights for some of the criteria. Thus, the combination of tests employed out of the total battery, as well as the specific weights, differs with the particular criterion. An example of such a classification battery is that developed by the Air Force for assignment of personnel to different training programs (DuBois, 1947). This battery, consisting of both paper-and-pencil and apparatus tests, provided stanine scores for pilots, navigators, bombardiers, and a few other air-crew specialties. By finding an individual's estimated criterion scores from the different regression equations, it was possible to predict whether, for example, he was better qualified as a pilot than as a navigator.

MAXIMIZING THE UTILIZATION OF TALENT. Differential prediction of criteria with a battery of tests permits a fuller utilization of available human resources than is possible with a single general test or with a composite score from a single regression equation. As we saw in the Taylor-Russell tables and elsewhere in this chapter, the effectiveness of any test in selecting personnel for a given job depends on the selection ratio. In classification decisions, we work with a smaller selection ratio and are thus able to assign better qualified persons to each job. If out of 100 applicants, 10 are needed to fill each of two different jobs, the selection ratio is 10 percent for each job, when separate predictors are employed for each. If a single predictor (such as a general intelligence test) were used to select applicants for both jobs, the selection ratio would be 20 percent, since we could do no better than take the top 20 applicants.

Even when predictors for the two jobs are highly correlated, so that some of the same applicants will qualify for both jobs, there is considerable gain from the use of separate predictors. This situation is illustrated in Table 18, which shows the mean standard criterion score of workers selected for each of two jobs by a selection strategy (single predictor) and by a classification strategy involving two different predictors, each validated against its own job criterion. If workers were assigned by chance, with no selection, the mean standard score in this scale would be zero. This would be the case if the selection ratio were 50 percent for each job, so that 100 percent of the applicants would have to be hired. Note that even under these conditions, job performance would be improved by the use of two predictors, as shown in the last row of the table. With two uncorrelated predictors, mean job performance would be .31 (approximately ⅓ of a standard deviation above the chance value). As the correlation between the predictors increases, the job effectiveness of

Table 18
Mean Standard Criterion Score of Persons Placed on Two Jobs by Selection or Classification Strategies

(Adapted from Brogden, 1951, p. 182)

Selection Ratio for Each Job	Selection: Single Predictor	Classification: Two predictors whose intercorrelation is				
		0	.20	.40	.60	.80
5%	.88	1.03	1.02	1.01	1.00	.96
10	.70	.87	.86	.84	.82	.79
20	.48	.68	.67	.65	.62	.59
30	.32	.55	.53	.50	.46	.43
40	.18	.42	.41	.37	.34	.29
50	.00	.31	.28	.25	.22	.17

the selected employees decreases; but it remains better than chance even when the correlation is .80. With lower selection ratios, we can of course obtain better qualified personnel. As can be seen in Table 18, however, for each selection ratio, mean job performance is better when applicants are chosen through classification than through selection strategies.

A practical illustration of the advantages of classification strategies is provided by the use of Aptitude Area scores in the assignment of personnel to military occupational specialties in the U.S. Army (Maier & Fuchs, 1972). Each Aptitude Area corresponds to a group of Army jobs requiring a similar pattern of aptitudes, knowledge, and interests. From a 13-test classification battery, combinations of three to five tests are used to find the individual's score in each Aptitude Area. Figure 20 shows the results of an investigation of 7,500 applicants for enlistment in which the use of Aptitude Area scores was compared with the use of a global screening test, the Armed Forces Qualification Test (AFQT). It will be noted that only 56 percent of this group reached or exceeded the 50th percentile on the AFQT, while 80 percent reached or exceeded the average standard score of 100 on their best Aptitude Area. Thus, when individuals are allocated to specific jobs on the basis of the aptitudes required for each job, a very large majority are able to perform as well as the average of the entire sample or better. This apparent impossibility, in which nearly everyone could be above average, can be attained by capitalizing on the fact that nearly everyone excels in *some* aptitude.

The same point is illustrated with a quite different population in a study of gifted children by Feldman and Bratton (1972). For demonstration purposes, 49 children in two fifth-grade classes were evaluated on

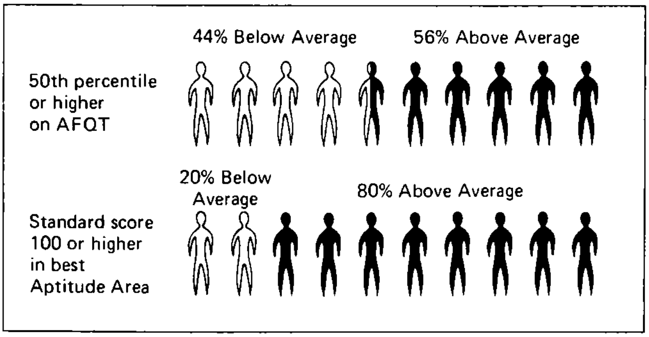


Fig. 20. Percentages Scoring Above Average on AFQT and on Best Aptitude Area of Army Classification Battery in a Sample of 7,500 Applicants for Enlistment.

(Data from U.S. Army Research Institute for the Behavioral and Social Sciences, Courtesy J. E. Uhlaner, 1974.)

each of 19 measures, all of which had previously been used to select students for special programs for the gifted. Among these measures were global scores on a group intelligence test and on an educational achievement battery, tests of separate aptitudes and separate academic areas such as reading and arithmetic, a test of creative thinking, grades in music and art, and teachers' nominations of the most "gifted" and the most "creative" children in each class. When the five highest ranking children selected by each criterion were identified, they included 92 percent of the group. Thus, it was again shown that nearly all members of a group will excel when multivariate criteria are employed.

DIFFERENTIAL VALIDITY. In the evaluation of a classification battery, the major consideration is its differential validity against the separate criteria. The object of such a battery is to predict the *difference* in each person's performance in two or more jobs, training programs, or other criterion situations. Tests chosen for such a battery should yield very different validity coefficients for the separate criteria. In a two-criterion classification problem, for example, the ideal test would have a high correlation with one criterion and a zero correlation (or preferably a negative correlation) with the other criterion. General intelligence tests are relatively poor for classification purposes, since they predict success about equally well in most areas. Hence, their correlations with the criteria to be differentiated would be too similar. An individual scoring high on such a test would be classified as successful for either assignment, and it would be impossible to predict in which he would do better. In a classification battery, we need some tests that are good predictors of criterion A and poor predictors of criterion B, and other tests that are poor predictors of A and good predictors of B.

Statistical procedures have been developed for selecting tests so as to maximize the differential validity of a classification battery (Brogden, 1951; Horst, 1954; Mollenkopf, 1950b; Thorndike, 1949). When the number of criteria is greater than two, however, the problem becomes very complex; no completely analytical solutions are yet available for these situations. In practice, various empirical approaches are followed to approximate the desired goals.

MULTIPLE DISCRIMINANT FUNCTIONS. An alternative way of handling classification decisions is by means of the multiple discriminant function (French, 1966). Essentially, this is a mathematical procedure for determining how closely the individual's scores on a whole set of tests approximate the scores typical of persons in a given occupation, curriculum, psychiatric syndrome, or other category. A person would then be assigned to the particular group he resembles most closely. Although the regression equation permits the prediction of degree of success in each field, the multiple discriminant function treats all persons in one category as of equal status. Group membership is the only criterion data utilized by this method. The discriminant function is useful when criterion scores are unavailable and only group membership can be ascertained. Some tests, for instance, are validated by administering them to persons in different occupations, although no measure of degree of vocational success is available for individuals within each field.

The discriminant function is also appropriate when there is a nonlinear relation between the criterion and one or more predictors. For example, in certain personality traits there may be an optimum range for a given occupation. Individuals having either more or less of the trait in question would thus be at a disadvantage. It seems reasonable to expect, for instance, that salesmen showing a moderately high amount of social dominance would be most likely to succeed, and that the chances of success would decline as scores move in either direction from this region. With the discriminant function, we would tend to select individuals falling within this optimum range. With the regression equation, on the other hand, the more dominant the score, the more favorable would be the predicted outcome. If the correlation between predictor and criterion were negative, of course, the regression equation would yield more favorable predictions for the low scorers. But there is no direct way whereby an intermediate score would receive maximum credit. Although in many instances the two techniques would lead to the same choices, there are situations in which persons would be differently classified by regression equations and discriminant functions. For most psychological testing purposes, regression equations provide a more effective technique. Under certain circumstances, however, the discriminant function is better suited to yield the required information.

STATISTICAL ANALYSES OF TEST BIAS

THE PROBLEM. If we want to use tests to predict outcomes in some future situation, such as an applicant's performance in college or on a job, we need tests with high predictive validity against the specific criterion. This requirement is commonly overlooked in the development of so-called culture-fair tests (to be discussed further in Ch. 12). In the effort to include in such tests only functions common to different cultures or subcultures, we may choose content that has little relevance to any criterion we wish to predict. A better solution is to choose criterion-relevant content and then investigate the possible effect of moderator variables on test scores. Validity coefficients, regression weights, and cutoff scores may vary as a function of differences in the examinees' experiential backgrounds. These values should therefore be checked within subgroups for whom there is reason to expect such effects.

It should be noted, however, that the predictive characteristics of test scores are less likely to vary among cultural groups when the test is intrinsically relevant to criterion performance. If a verbal test is employed to predict nonverbal job performance, a fortuitous validity may be found in one cultural group because of traditional associations of past experiences within that culture. In a group with a different experiential background, however, the validity of the test may disappear. On the other hand, a test that directly samples criterion behavior, or one that measures essential prerequisite skills, is likely to retain its validity in different groups.

Since the mid-1960s, there has been a rapid accumulation of research on possible ethnic differences in the predictive meaning of test scores. In this connection, the EEOC *Guidelines* (Appendix B) explicitly state: "Data must be generated and results separately reported for minority and nonminority groups wherever technically feasible." The functions and implications of separate validation studies are also discussed in the reports of the APA task forces on the testing of minority groups for both educational and employment purposes (American Psychological Association, 1968; Cleary, Humphreys, Kendrick, & Wesman, 1975). The large majority of studies conducted thus far have dealt with black Americans, although a few have included other ethnic minorities. The problems investigated are generally subsumed under the heading of *test bias*. In this context, the term "bias" is employed in its well-established statistical sense, to designate constant or systematic error as opposed to chance error. This is the same sense in which we speak of a biased sample, in contrast to a random sample. The principal questions that have been raised regarding test bias pertain to validity coefficients (slope bias) and to the relationship between group means on the test and on the criterion (intercept bias). These questions will be examined in the next two sections.

SLOPE BIAS. To facilitate an understanding of the technical aspects of test bias, let us begin with a scatter diagram, or bivariate distribution, such as those illustrated in Chapter 5 (Figs. 8, 9, 10—see especially Fig. 10, p. 110). For the present purpose, the horizontal axis (X) represents scores on a test and the vertical axis (Y) represents criterion scores, such as college grade-point average or an index of job performance. It will be recalled that the tally marks, showing the position of each individual on both test and criterion, indicate the direction and general magnitude of the correlation between the two variables. The line of best fit drawn through these tally marks is known as the regression line, and its equation is the regression equation. In this example, the regression equation would have only one predictor. The multiple regression equations discussed earlier in this chapter have several predictors, but the principle is the same.

When both test and criterion scores are expressed as standard scores ($SD = 1.00$), the *slope* of the regression line equals the correlation coefficient. For this reason, if a test yields a significantly different validity coefficient in the two groups, this difference is described as slope bias. Figure 21 provides schematic illustrations of regression lines for several

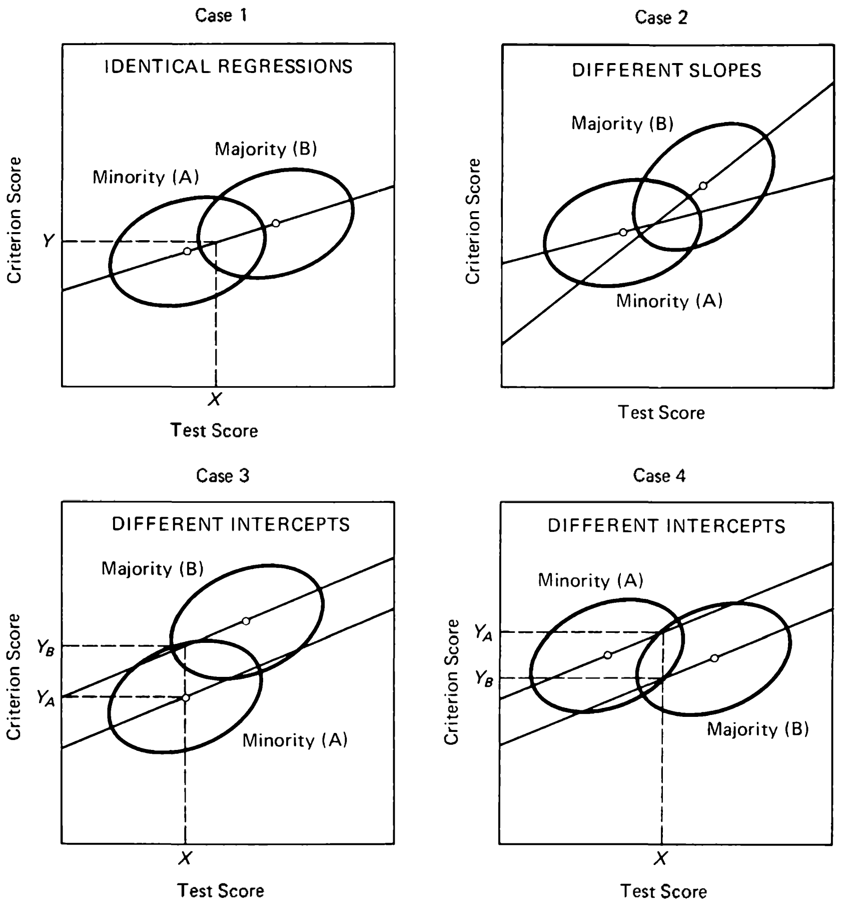

Case 1
IDENTICAL REGRESSIONS
Majority (B)
Minority (A)
Y
X
Criterion Score
Test Score
Case 2
DIFFERENT SLOPES
Majority (B)
Minority (A)
Criterion Score
Test Score
Case 3
DIFFERENT INTERCEPTS
Majority (B)
Minority (A)
Y_B
Y_A
X
Criterion Score
Test Score
Case 4
DIFFERENT INTERCEPTS
Minority (A)
Majority (B)
Y_A
Y_B
X
Criterion Score
Test Score

FIG. 21. Slope and Intercept Bias in Predicting Criterion Scores. The ellipses show the regions within which members of each group fall when their test scores are plotted against their criterion performance.

(Cases 1, 2, and 3 adapted from M. A. Gordon, 1953, p. 3.)

bivariate distributions. The ellipses represent the region within which the tally marks of each sample would fall. Case 1 shows the bivariate distribution of two groups with different means in the predictor, but with identical regression lines between predictor and criterion. In this case, there is no test bias, since any given test score (X) corresponds to the identical criterion score (Y) in both groups. Case 2 illustrates slope bias, with a lower validity coefficient in the minority group.

In comparative validation studies, certain methodological precautions must be observed. For example, the use of ratings as a criterion in this type of study may yield results that are out of line with those obtained with more objective criteria (Bass & Turner, 1973; Campbell, Crooks, Mahoney, & Rock, 1973; Kirkpatrick, Ewen, Barrett, & Katzell, 1968). A second example involves the comparison of ethnic samples from different institutions. In such cases, ethnic and institutional factors are likely to be confounded in the results (Kirkpatrick et al., 1969).

A common difficulty arises from the fact that in several studies the number of cases in the minority sample was much smaller than in the majority sample. Under these conditions, the same validity coefficient could be statistically significant in the majority sample and not significant in the minority sample. With 100 cases, for example, a correlation of .27 is clearly significant at the .01 level; with 30 cases, the same correlation falls far short of the minimum value required for significance even at the .05 level. For this reason, the proper procedure to follow in such differential validation studies is to evaluate the difference between the two validity coefficients, rather than testing each separately for significance (Humphreys, 1973; *Standards,* 1974, E9). By the latter procedure, one could easily "demonstrate" that a test is valid for, let us say, whites and not valid for blacks. All that would be needed for this purpose is a large enough group of whites and a small enough group of blacks! Cross-validation of results in a second pair of independent samples is also desirable, to check whether the group yielding higher validity in the first study still does so in the second.

A sophisticated statistical analysis of the results of 19 published studies reporting validity coefficients for black and white employment samples casts serious doubt on the conclusions reached in some of the earlier studies (Schmidt, Berner, & Hunter, 1973). Taking into account the obtained validities and the size of samples in each study, the investigators demonstrated that the discrepancies in validity coefficients found between blacks and whites did not differ from chance expectancy. The same conclusion was reached when only studies including approximately equal numbers of blacks and whites were surveyed (O'Connor, Wexley, & Alexander, 1975). Well-designed recent studies corroborate these findings for industrial samples (Campbell et al., 1973) and army personnel (Maier & Fuchs, 1973). In their review of validity studies with minority groups in the chapter on Personnel Psychology for the *Annual Review of Psychology,* Bray and Moses (1972, p. 554) concluded: "It does appear, however, that the closer the study comes to the ideal, the less likelihood there is of finding differential validity."

Similar results have been obtained in investigations of black and white college students. Validity coefficients of the College Board Scholastic Aptitude Test for black students were generally as high as those obtained for white students, or higher. These relationships were found when the black and white samples were attending the same colleges, as well as when they were attending separate colleges (Cleary, 1968; Kendrick & Thomas, 1970; Stanley & Porter, 1967). Working at a very different level, Mitchell (1967) studied the validities of two educational readiness tests against end-of-year achievement test scores of first-grade schoolchildren. In the large samples of black and white children tested, validities of total scores and of subtests were very similar for the two ethnic groups, although tending to run somewhat higher for the blacks. These findings were consistently corroborated in later studies of black and white children, as well as in comparisons of children classified by father's occupational level and by indices of educational and socioeconomic level of the community.[6]

INTERCEPT BIAS. Even when a test yields the same validity coefficients for two groups, it may show intercept bias. The intercept of a regression line refers to the point at which the line intersects the axis. A test exhibits intercept bias if it systematically underpredicts or overpredicts criterion performance for a particular group. Let us look again at Case 1 in Figure 21, in which majority and minority samples show identical regressions. Under these conditions, there is neither slope nor intercept bias. Although the groups differ significantly in mean test scores, they show a corresponding difference in criterion performance. In Case 3, on the other hand, the two groups have regression lines with the same slope but different intercepts. In this case, the majority group (B) has a higher intercept than the minority group (A), that is, the majority regression line intersects the Y axis at a higher point than does the minority regression line. Although the validity coefficients computed within each group are equal, any test score (X) will correspond to different criterion scores in the two groups, as shown by points Y_A and Y_B. The same test score thus has a different predictive meaning for the two groups. In this situation, the majority mean again exceeds the minority mean in both test and criterion, as it did in Case 1. Because of the intercept difference, however, the use of the majority regression line for both groups would *overpredict* the criterion performance of minority group members. If a single cutoff score (X) were applied to both groups, it would discriminate in favor of the minority group. Intercept bias discriminates against the group with the higher intercept.

---

[6] Summaries of these studies can be obtained from the test publisher, Harcourt Brace Jovanovich, Inc.

Psychologists who are concerned about the possible unfairness of tests for minority group members visualize a situation illustrated in Case 4. Note that, in this case, the majority excels on the test, but majority and minority perform equally well on the criterion. The minority group now has the higher intercept. Selecting all applicants in terms of a test cutoff established for the majority group would thus discriminate unfairly against the minority. Under these conditions, use of the majority regression line for both groups *underpredicts* the criterion performance of minority group members. This situation is likely to occur when a large proportion of test variance is irrelevant to criterion performance and measures functions in which the majority excels the minority. A thorough job analysis and satisfactory test validity provide safeguards against the choice of such a test.

Reilly (1973) has demonstrated mathematically that Case 3 will occur if the two groups differ in a third variable (e.g., sociocultural background) which correlates positively with both test and criterion. Under these conditions, the test overpredicts the performance of minority group members and the use of the same cutoff scores for both groups favors the minority. The findings of empirical studies do in fact support this expectation. Using principally a statistical procedure developed by Gulliksen and Wilks (1950), investigators have checked intercept bias in the prediction of college grades (Cleary, 1968; Temp, 1971), law school grades (Linn, 1975), performance in army and air force training programs (M. A. Gordon, 1953; Maier & Fuchs, 1973; Shore & Marion, 1972), and a wide variety of industrial criteria (Campbell et al., 1973; Grant & Bray, 1970; Ruch, 1972).

It is interesting to note that the same results have been obtained when comparisons were made between groups classified according to educational or socioeconomic level. The Army Classification Battery tended to overpredict the performance of high school dropouts and underpredict the performance of college graduates in training programs for military occupational specialties (Maier, 1972). Similarly, the college grades of students whose fathers were in the professions were underpredicted from various academic aptitude tests, while the grades of students from lower occupational levels tended to be overpredicted (Hewer, 1965). In all these studies, comparisons of higher-scoring and lower-scoring groups revealed either no significant difference in intercepts or a slight bias in favor of the groups scoring lower on the tests.

The problem of test bias is complex—more so than appears in this simplified exposition. This is an area in which the statistically unsophisticated must step warily. Some investigators have proposed alternative definitions of test bias, not in terms of predicted criterion scores, but in terms of percentage of minority and majority group members who exceed cutoff points on test and criterion (Cole, 1972; Linn, 1973, 1975; Schmidt & Hunter, 1974; Thorndike, 1971). In a comparative evaluation of several alternative models of test bias, however, other psychometricians (Gross & Su, 1975; Petersen, 1974; Petersen & Novick, 1976) questioned the conceptual and methodological soundness of these procedures and formulated a comprehensive mathematical model for culture-fair personnel selection. Based on decision theory, this model combines data on probability of different outcomes with judgments of the relative utility of each outcome, such as accepting an unsuccessful applicant or rejecting a potential success. It is argued that there is no best single model for fair selection. The proposed model provides a means of formulating, within each context, a decision strategy that maximizes the expected overall utility in accordance with specified testing goals and judged utilities of different outcomes.

Research on test bias is still being actively pursued, with regard to both statistical theory and empirical investigations. In our present state of knowledge, there is insufficient basis for the use of different cutoff scores for different subgroups of the population. Statistical adjustments in test scores, cutoffs, and prediction formulas hold little promise as a means of correcting social inequities. More constructive solutions are suggested by other approaches discussed earlier in this chapter. One is illustrated by multiple aptitude testing and classification strategies, which permit the fullest utilization of the diverse aptitude patterns fostered by different cultural backgrounds. Another approach is through adaptive treatments, such as individualized training programs. In order to maximize the fit of such programs to individual characteristics, it is essential that tests reveal as fully and accurately as possible the person's present level of development in the requisite abilities.

CHAPTER 8

Item Analysis

FAMILIARITY with the basic concepts and techniques of item analysis, like knowledge about other phases of test construction, can help the test user in his evaluation of published tests. In addition, item analysis is particularly relevant to the construction of informal, local tests, such as the quizzes and examinations prepared by teachers for classroom use. Some of the general guidelines for effective item writing, as well as the simpler statistical techniques of item analysis, can materially improve classroom tests and are worth using even with small groups.

Items can be analyzed qualitatively, in terms of their content and form, and quantitatively, in terms of their statistical properties. Qualitative analysis includes the consideration of content validity, discussed in Chapter 6, and the evaluation of items in terms of effective item-writing procedures, to be discussed in Chapter 14. Quantitative analysis includes principally the measurement of item difficulty and item validity. Both the validity and the reliability of any test depend ultimately on the characteristics of its items. High reliability and validity can be built into a test in advance through item analysis. Tests can be improved through the selection, substitution, or revision of items.

Item analysis makes it possible to shorten a test, and at the same time increase its validity and reliability. Other things being equal, a longer test is more valid and reliable than a shorter one. The effect of lengthening or shortening a test on the reliability coefficient was discussed in Chapter 5, where the Spearman-Brown formula for estimating this effect was also presented. These estimated changes in reliability occur when the discarded items are equivalent to those that remain, or when equivalent new items are added to the test. Similar changes in validity will result from the deletion or addition of items of equivalent validity. All such estimates of change in reliability or validity refer to the lengthening or shortening of tests through a *random* selection of items, without item analysis. When a test is shortened by eliminating the least satisfactory items, however, the short test may be more valid and reliable than the original, longer instrument.

ITEM DIFFICULTY

PERCENTAGE PASSING. For most testing purposes, the difficulty of an item is defined in terms of the percentage of persons who answer it correctly. The easier the item, the larger will this percentage be. A word that is correctly defined by 70 percent of the standardization sample ($p = .70$) is regarded as easier than one that is correctly defined by only 15 percent ($p = .15$). It is customary to arrange items in order of difficulty, so that examinees begin with relatively easy items and proceed to items of increasing difficulty. This arrangement gives the individual confidence in approaching the test and also reduces the likelihood of his wasting much time on items beyond his ability to the neglect of easier items he can correctly complete.

In the process of test construction, a major reason for measuring item difficulty is to choose items of suitable difficulty level. Most standardized ability tests are designed to assess as accurately as possible each individual's level of attainment in the particular ability. For this purpose, if no one passes an item, it is excess baggage in the test. The same is true of items that everyone passes. Neither of these types of items provides any information about individual differences. Since such items do not affect the variability of test scores, they contribute nothing to the reliability or validity of the test. The closer the difficulty of an item approaches 1.00 or 0, the less differential information about examinees it contributes. Conversely, the closer the difficulty level approaches .50, the more differentiations the item can make. Suppose out of 100 persons, 50 pass an item and 50 fail it ($p = .50$). This item enables us to differentiate between each of those who passed it and each of those who failed it. We thus have 50 × 50 or 2500 paired comparisons, or bits of differential information. An item passed by 70 percent of the persons provides 70 × 30 or 2100 bits of information; one passed by 90 percent provides 90 × 10 or 900; one passed by 100 percent provides 100 × 0 or 0. The same relationships would hold for harder items, passed by fewer than 50 percent.

For maximum differentiation, then, it would seem that one should choose all items at the .50 difficulty level. The decision is complicated, however, by the fact that items within a test tend to be intercorrelated. The more homogeneous the test, the higher will these intercorrelations be. In an extreme case, if all items were perfectly intercorrelated and all were of .50 difficulty level, the same 50 persons out of 100 would pass each item. Consequently, half of the examinees would obtain perfect scores and the other half zero scores. Because of item intercorrelations, it is best to select items with a moderate spread of difficulty level, but whose average difficulty is .50.

INTERVAL SCALES. The percentage of persons passing an item expresses item difficulty in terms of an ordinal scale; that is, it correctly indicates the rank order or relative difficulty of items. For example, if Items 1, 2, and 3 are passed by 30, 20, and 10 percent of the cases, respectively, we can conclude that Item 1 is the easiest and Item 3 is the hardest of the three. But we cannot infer that the difference in difficulty between Items 1 and 2 is equal to that between Items 2 and 3. Equal percentage differences would correspond to equal differences in difficulty only in a rectangular distribution, in which the cases were uniformly distributed throughout the range. This problem is similar to that encountered in connection with percentile scores, which are also based on percentages of cases. It will be recalled from Chapter 4 that percentile scores do not represent equal units, but differ in magnitude from the center to the extremes of the distribution (Fig. 4, Ch. 4).

If we assume a normal distribution of the trait measured by any given item, the difficulty level of the item can be expressed in terms of an equal-unit interval scale by reference to a table of normal curve frequencies. In Chapter 4 we saw, for example, that approximately 34 percent of the cases in a normal distribution fall between the mean and a distance of 1σ in either direction (Fig. 3, Ch. 4). With this information, we can examine Figure 22, which shows the difficulty level of an item passed by 84 percent of the cases. Since it is the persons in the upper part of the distribution who pass and those in the lower part who fail, this 84 percent includes the upper half (50%) plus 34 percent of the cases from the lower half ($50 + 34 = 84$). Hence, the item falls 1σ *below* the mean, as shown in Figure 22. An item passed by 16 percent of the cases would fall 1σ *above* the mean, since above this point there are 16 percent of the cases

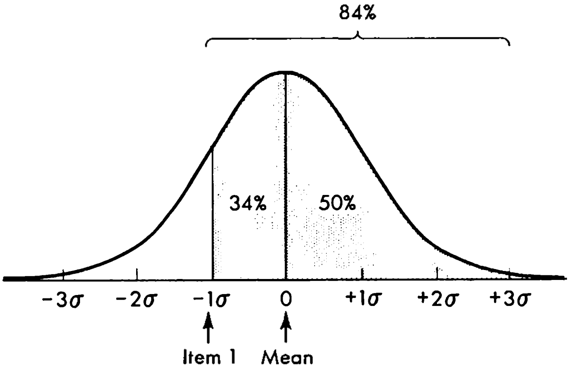


FIG. 22. Relation between Percentage of Persons Passing an Item and the Item Difficulty in Normal Curve Units.

$(50 - 34 = 16)$. An item passed by exactly 50 percent of the cases falls at the mean and would thus have a 0 value on this scale. The more difficult items have plus values, the easier items minus values. The difficulty value corresponding to any percentage passing can be found by reference to a normal curve frequency table, given in any standard statistics text.

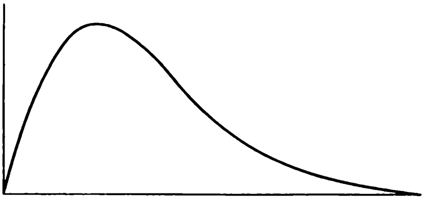
Because item difficulties expressed in terms of normal curve σ-units involve negative values and decimals, they are usually converted into a more manageable scale. One such scale, employed by Educational Testing Service in its test development, uses a unit designated by the Greek letter *delta* (Δ). The relation between Δ and normal curve σ-values (x) is shown below:

$$\Delta = 13 + 4x$$

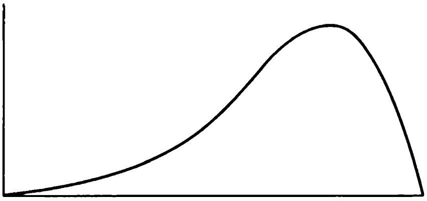
The constants 13 and 4 were chosen arbitrarily in order to provide a scale that eliminates negative values and yields a range of integers wide enough to permit the dropping of decimals. An item passed by nearly 100 percent of the cases (99.87%), falling at -3σ, would have a Δ of: $13 + (4)(-3) = 1$. This is the lowest value likely to be found in most groups. At the other extreme, an item passed by less than 1 percent (0.13%) of the cases, will have a value of $+3\sigma$ and a Δ of: $13 + (4)(3) = 25$. An item falling at the mean will have a 0 σ-value and a Δ of: $13 + (4)(0) = 13$. The Δ scale is thus a scale in which practically all items will fall between 1 and 25, and the mean difficulty value within any given group corresponds to 13.

An important practical advantage of the Δ scale over other possible conversions is that a table is available (Fan, 1952) from which Δ can be found by simply entering the value of p (proportion of persons passing the item). The table eliminates the necessity of looking up normal curve σ-values and transforming these values to Δ's. For most practical purposes, an ordinal measure of item difficulty, such as percentage passing, is adequate. For more precise statistical analyses, requiring the measurement of difficulty on an interval scale, Δ values can be obtained with little additional effort.

DISTRIBUTION OF TEST SCORES. The difficulty of the test as a whole is, of course, directly dependent on the difficulty of the items that make up the test. A comprehensive check of the difficulty of the total test for the population for which it is designed is provided by the distribution of total scores. If the standardization sample is a representative cross section of such a population, then it is generally expected that the scores will fall roughly into a normal distribution curve.



A. Piling at Lower End of the Scale



B. Piling at Upper End of the Scale

FIG. 23. Skewed Distribution Curves.

Let us suppose, however, that the obtained distribution curve is not normal but clearly skewed, as illustrated in Figure 23, Parts A and B. The first of these distributions, with a piling of scores at the low end, suggests that the test has too high a floor for the group under consideration, lacking a sufficient number of easy items to discriminate properly at the lower end of the range. The result is that persons who would normally scatter over a considerable range obtain zero or near-zero scores on this test. A peak at the low end of the scale is therefore obtained. This artificial piling of scores is illustrated schematically in Figure 24, in which a normally distributed group yields a skewed distribution on a particular test. The opposite skewness is illustrated in Part B of Figure 23, with the scores piled up at the upper end, a finding that suggests insufficient test ceiling. Administering a test designed for the general population to selected samples of college or graduate students will usually yield such a skewed distribution, a number of students obtaining nearly perfect scores. With such a test, it is impossible to measure individual differences among the more able subjects in the group. If more difficult items had been included in the test, some individuals would undoubtedly have scored higher than the present test permits.

When the standardization sample yields a markedly nonnormal distribution on a test, the difficulty level of the test is ordinarily modified until a normal curve is approximated. Depending on the type of deviation from normality that appears, easier or more difficult items may be added, other items eliminated or modified, the position of items in the scale altered, or the scoring weights assigned to certain responses revised. Such adjustments are continued until the distribution becomes at least roughly normal. Under these conditions, the most likely score, obtained by the largest number of subjects, usually corresponds to about 50 percent correct items. To the layman who is unfamiliar with the methods of psychological test construction, a 50 percent score may seem shockingly low. It is sometimes objected, on this basis, that the examiner has set too low a standard of passing on the test. Or the inference is drawn that the group tested is a particularly poor one. Both conclusions, of course, are totally meaningless when viewed in the light of the procedures followed in developing psychological tests. Such tests are deliberately constructed and specifically modified so as to yield a mean score of approximately 50 percent correct. Only in this way can the maximum differentiation between individuals at all ability levels be obtained with the test. With a mean of approximately 50 percent correct items, there is the maximum opportunity for a normal distribution, with individual scores spreading widely at both extremes.[1]

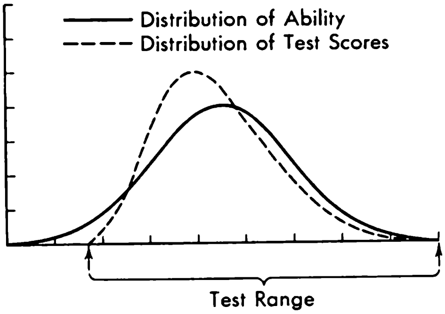


Fig. 24. Skewness Resulting from Insufficient Test Floor.

The difficulty level of the items that constitute a test determines not only the mean difficulty level of the test and the test floor and ceiling, but also the spread of test scores. As mentioned earlier, the maximum spread of total test scores is obtained with items whose difficulty levels cluster closely around .50. The fact that such a selection of items yields wider differentiation than would be obtained with items that are themselves widely distributed in difficulty level is illustrated in Figure 25. The three distributions of total scores given in Figure 25 were obtained by Ebel (1965) with three 16-item tests assembled for this purpose. The items in Test 1 were chosen so as to cluster close to the .50 difficulty level; those in Test 2 are widely distributed over the entire difficulty range; and those in Test 3 fall at the two extremes of difficulty. It will be noted that the widest spread of total test scores was obtained with the items

---

[1] Actually, the normal curve provides finer discrimination at the ends than at the middle of the scale. Equal discrimination at all points of the scale would require a rectangular distribution. The normal curve, however, has an advantage if subsequent statistical analyses of scores are to be conducted, because many current statistical techniques assume approximate normality of distribution. For this and other reasons, it is likely that most tests designed for general use will continue to follow a normal-curve pattern for some time to come.

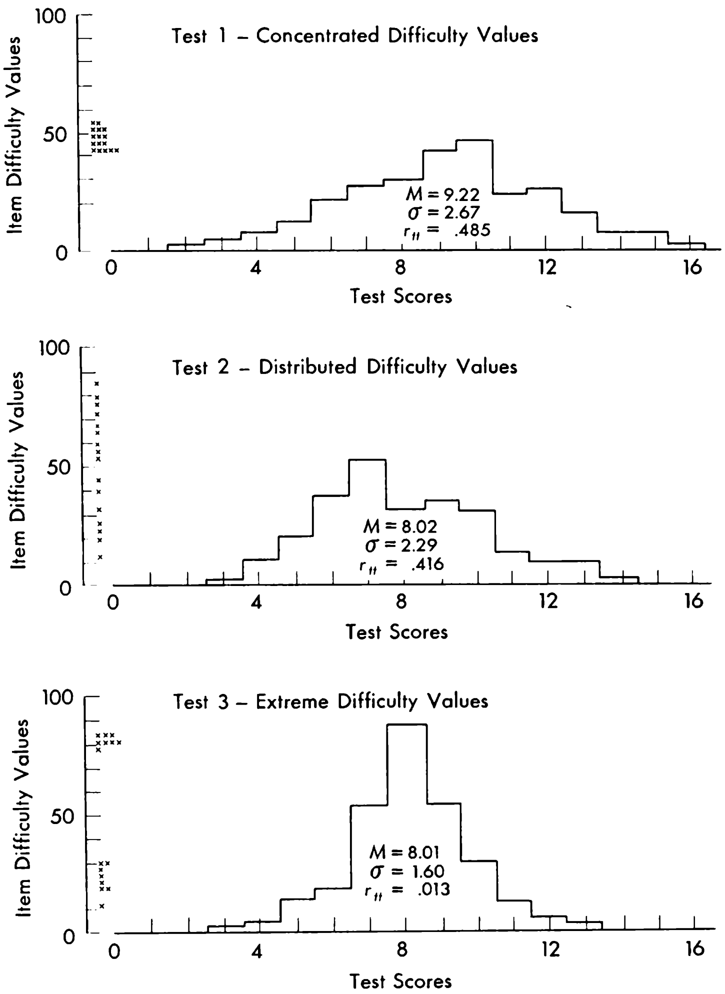


Fig. 25. Relation between Distribution of Test Scores and Distribution of Item Difficulty Values.

(From Ebel, 1965, p. 363.)

concentrated around .50. The reliability coefficient was also highest in this case, and it was particularly low in the test composed of items with extreme difficulty values (Test 3). This simple demonstration serves only to clarify the point; similar conclusions have been reached in more technical analyses of the problem, with both statistical and experimental procedures (Cronbach & Warrington, 1952; Lord, 1952; Lord & Novick, 1968).

RELATING ITEM DIFFICULTY TO TESTING PURPOSE. Standardized psychological tests have generally been designed to elicit maximum differentiation among individuals at all levels. Our discussion of item difficulty thus far has been directed to this type of test. In the construction of tests to serve special purposes, however, the choice of appropriate item difficulties, as well as the optimal form of the distribution of test scores, depends upon the type of discrimination sought. Accordingly, tests designed for screening purposes should utilize items whose difficulty values come closest to the desired selection ratio (Lord, 1953). For example, to select the upper 20 percent of the cases, the best items are those clustering around a p of .20. Since in a screening test no differentiation is required *within* the accepted or rejected groups, the most effective use of testing time is obtained when items cluster near the critical cutoff. It follows, for instance, that if a test is to be used to screen scholarship applicants from a college population, the items should be considerably more difficult than the average for that population. Similarly, if slow learners are being selected for a remedial training program, items that are much easier than average would be desirable.

Another illustration is provided by the National Assessment of Educational Progress (Womer, 1970). Designed as an attempt to obtain direct information on the outcomes of education throughout the United States, this project examined carefully chosen representative samples of the population at four age levels: 9-, 13-, and 17-year-olds and young adults between the ages of 26 and 35. The project was not concerned with the achievement of individuals. Its aim was to describe the knowledge, understanding, and skills manifested by the American population at the specified age levels. Within each of the content areas and each of the age groups tested, achievement was assessed in terms of three questions: (1) What do *almost all* persons know? (2) What does the *typical or average* person know? (3) What do the *most able* persons know? To answer these questions, exercises[2] were prepared at three difficulty levels: one third of the exercises were "easy" (.90); one third were "average" (.50); and one third were "difficult" (.10). The actual percentages of persons passing each exercise varied somewhat around these values. But the goal of the test constructors was to approximate the three values as closely as possible.

---

[2] Because of the nature of many of the tests, the term "exercise" was considered more appropriate than "item." For purposes of the present discussion, they can be regarded as items.

A third example of the choice of item difficulty levels in terms of special testing goals is to be found in mastery testing. It will be recalled (Ch. 4) that mastery testing is often associated with criterion-referenced testing. If the purpose of the test is to ascertain whether an individual has adequately mastered the basic essentials of a skill or whether he has acquired the prerequisite knowledge to advance to the next step in a learning program, then the items should probably be at the .80 or .90 difficulty level. Under these conditions, we would expect the majority of those taking the examination to complete nearly all items correctly. Thus, the very easy items (even those passed by 100% of the cases), which are discarded as nondiscriminative in the usual standardized test, are the very items that would be included in a mastery test. Similarly, a pretest, administered prior to a learning unit to determine whether any of the students have already acquired the skills to be taught, will yield very low percentages of passing for each item. In this case, items with very low or even zero *p* values should not be discarded, since they reveal what remains to be learned.

It is apparent from these examples that the appropriate difficulty level of items depends upon the purpose of the test. Although in most testing situations items clustering around a medium difficulty (.50) yield the maximum information about the individual's performance level, decisions about item difficulty cannot be made routinely, without knowing how the test scores will be used.

ITEM VALIDITY

ITEM-CRITERION RELATIONSHIPS. All indices of item validity are based on the relationship between item response and criterion performance. Any criterion employed for test validation is also suitable for item validation. Item analysis can be employed to improve not only the convergent but also the discriminant validity of a test (see Ch. 6). Thus, items can be chosen on the basis of high correlation with a criterion and low correlation with any irrelevant variable that may affect test performance. In the development of an arithmetic reasoning test, for example, items that correlate significantly with a reading comprehension test would be discarded.

Since item responses are generally recorded as pass or fail, the measurement of item validity usually involves a dichotomous variable (the item) and a continuous variable (the criterion). In certain situations, the criterion too may be dichotomous, as in graduation versus nongraduation from college or success versus failure on a job. Moreover, a continuous criterion may be dichotomized for purposes of analysis. The basic relationship between item and criterion is illustrated by the three item characteristic curves reproduced in Figure 26. Using fictitious data, each of these curves shows the percentages of persons in each class-interval of criterion score who pass the item. It can be seen that Item 1 has a low validity, since it is passed by nearly the same proportion of persons throughout the criterion range. Items 2 and 3 are better, showing a closer correspondence between percentage passing and criterion score. Of these two, 3 is the more valid item, since its curve rises more steeply.

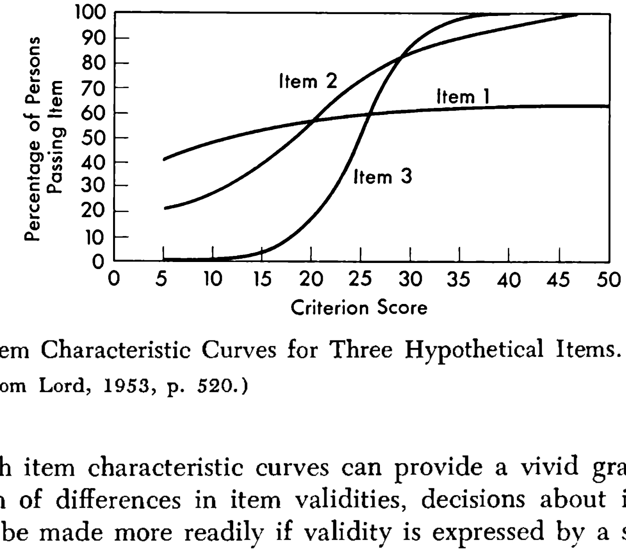


FIG. 26. Item Characteristic Curves for Three Hypothetical Items. (Adapted from Lord, 1953, p. 520.)

Although item characteristic curves can provide a vivid graphic representation of differences in item validities, decisions about individual items can be made more readily if validity is expressed by a single numerical index for each item. Over fifty different indices of item validity have been developed and used in test construction. One difference among them pertains to their applicability to dichotomous or continuous measures. Among those applicable to dichotomous variables, moreover, some assume a continuous and normal distribution of the underlying trait, on which the dichotomy has been artificially imposed; others assume a true dichotomy. Another difference concerns the relation of item difficulty to item validity. Certain indices measure item validity independently of item difficulty. Others yield higher validities for items close to the .50 difficulty level than for those at the extremes of difficulty.

Despite differences in procedure and assumptions, most item validity indices provide closely similar results. Although the numerical values of the indices may differ, the items that are retained and those that are rejected on the basis of different validity indices are largely the same. In fact, the variation in item validity data from sample to sample is generally greater than that among different methods. For this reason, the choice of method is often based on the amount of computational labor required and the availability of special computational aids. Among the published computational aids are a number of abacs or nomographs. These are computing diagrams with which, for example, the value of an item-criterion correlation can be read directly if the percentages of persons passing the item in high and low criterion groups are known (Guilford & Fruchter, 1973, pp. 445–458; Henrysson, 1971).

USE OF EXTREME GROUPS. A common practice in item analysis is to compare the proportion of cases who pass an item in contrasting criterion groups. When the criterion is measured along a continuous scale, as in the case of course grades, job ratings, or output records, upper (U) and lower (L) criterion groups are selected from the extremes of the distribution. Obviously, the more extreme the groups the sharper will be the differentiation. But the use of very extreme groups, such as upper and lower 10 percent, would reduce the reliability of the results because of the small number of cases utilized. In a normal distribution, the optimum point at which these two conditions balance is reached with the upper and lower 27 percent (Kelley, 1939). When the distribution is flatter than the normal curve, the optimum percentage is slightly greater than 27 and approaches 33 (Cureton, 1957b). With small groups, as in an ordinary classroom, the sampling error of item statistics is so large that only rough results can be obtained. Under these conditions, therefore, we need not be too concerned about the exact percentage of cases in the two contrasted groups. Any convenient number between 25 percent and 33 percent will serve satisfactorily.

With the large and normally distributed samples employed in the development of standardized tests, it has been customary to work with the upper and lower 27 percent of the criterion distribution. Many of the tables and abacs prepared to facilitate the computation of item validity indices are based on the assumption that the "27 percent rule" has been followed. As high-speed computers become more generally available, it is likely that the various labor-saving procedures developed to facilitate item analysis will be gradually replaced by more exact and sophisticated methods. With computer facilities, it is better to analyze the results for the entire sample, rather than working with upper and lower extremes. Mathematical procedures have also been developed for measuring item validity from the item characteristic curves, but their application is not feasible without access to a computer (Baker, 1971; Henrysson, 1971; Lord & Novick, 1968).

SIMPLE ANALYSIS WITH SMALL GROUPS. Because item analysis is frequently conducted with small groups, such as the students who have taken a classroom quiz, we shall consider first a simple procedure especially suitable for this situation. Let us suppose that in a class of 60 students we have chosen the 20 students (33%) with the highest and the 20 with the lowest scores. We now have three groups of papers which we may call the Upper (U), Middle (M), and Lower (L) groups. First we need to tally the correct responses to each item given by students in the three groups. This can be done most readily if we list the item numbers in one column and prepare three other columns headed U, M, and L. As we come to each student's paper, we simply place a tally next to each item he answered correctly. This is done for each of the 20 papers in the U group, then for each of the 20 in the M group, and finally for each of the 20 in the L group. We are now ready to count up the tallies and record totals for each group as shown in Table 19. For illustrative purposes, the first seven items have been entered. A rough index of the validity or discriminative value of each item can be found by subtracting the number of persons answering it correctly in the L group from the number answering it correctly in the U group. These U-L differences are given in the last column of Table 19. A measure of item difficulty can be obtained with the same data by adding the number passing each item in all three criterion groups (U + M + L).

TABLE 19
Simple Item Analysis Procedure: Number of Persons Giving Correct Response in Each Criterion Group

Item	U (20)	M (20)	L (20)	Difficulty (U + M + L)	Discrimination (U − L)
1	15	9	7	31	8
2	20	20	16	56*	4
3	19	18	9	46	10
4	10	11	16	37	− 6*
5	11	13	11	35	0*
6	16	14	9	39	7
7	5	0	0	5*	5
.					
.					
.					
.					
75					

* Items chosen for discussion.

Examination of Table 19 reveals four questionable items that have been identified for further consideration or for class discussion. Two items, 2 and 7, have been singled out because one seems to be too easy, having been passed by 56 out of 60 students, and the other too difficult, having been passed by only 5. Items 4 and 5, while satisfactory with regard to difficulty level, show a negative and zero discriminative value, respectively. We would also consider in this category any items with a very small positive U–L difference, of roughly three or less when groups of approximately this size are being compared. With larger groups, we would expect larger differences to occur by chance in a nondiscriminating item.

The purpose of item analysis in a teacher-made test is to identify deficiencies either in the test or in the teaching. Discussing questionable items with the class is often sufficient to diagnose the problem. If the wording of the item was at fault, it can be revised or discarded in subsequent testing. Discussion may show, however, that the item was satisfactory, but the point being tested had not been properly understood. In that case, the topic may be reviewed and clarified. In narrowing down the source of the difficulty, it is often helpful to carry out a supplementary analysis, as shown in Table 20, with at least some of the items

Table 20
Response Analysis of Individual Items

Item	Group	Response Options[a]				
		1	2	3	4	5
2	Upper	0	0	0	*20*	0
	Lower	2	0	1	*16*	1
4	Upper	0	*10*	9	0	1
	Lower	2	*16*	2	0	0
5	Upper	2	3	2	*11*	2
	Lower	1	3	3	*11*	2
7	Upper	*5*	3	5	4	3
	Lower	*0*	5	8	3	4
.						
.						
.						

[a] Correct options have been italicized.

chosen for discussion. This tabulation gives the number of students in the U and L groups who chose each option in answering the particular items.

Although Item 2 has been included in Table 20, there is little more we can learn about it by tabulating the frequency of each wrong option, since only 4 persons in the L group and none in the U group chose wrong answers. Discussion of the item with the students, however, may help to determine whether the item as a whole was too easy and therefore of little intrinsic value, whether some defect in its construction served to give away the right answer, or whether it is a good item dealing with a point that happened to have been effectively taught and well remembered. In the first case, the item would probably be discarded, in the second it would be revised, and in the third it would be retained unchanged.

The data on Item 4 suggest that the third option had some unsuspected implications that led 9 of the better students to prefer it to the correct alternative. The point could easily be settled by asking those students to explain why they chose it. In Item 5, the fault seems to lie in the wording either of the stem or of the correct alternative, because the students who missed the item were uniformly distributed over the four wrong options. Item 7 is an unusually difficult one, which was answered incorrectly by 15 of the U and all of the L group. The slight clustering of responses on incorrect option 3 suggests a superficial attractiveness of this option, especially for the more easily misled L group. Similarly, the lack of choices of the correct response (option 1) by any of the L group suggests that this alternative was so worded that superficially, or to the uninformed, it seemed wrong. Both of these features, of course, are desiderata of good test items. Class discussion might show that Item 7 is a good item dealing with a point that few class members had actually learned.

THE INDEX OF DISCRIMINATION. If the numbers of persons passing each item in U and L criterion groups are expressed as percentages, the difference between these two percentages provides an index of item validity that can be interpreted independently of the size of the particular sample in which it was obtained. This index has been repeatedly described in the psychometric literature (see, e.g., Ebel, 1965; Johnson, 1951; Mosier & McQuitty, 1940) and has been variously designated as *ULI, ULD,* or simply *D*. Despite its simplicity, it has been shown to agree quite closely with other, more elaborate measures of item validity (Engelhart, 1965). The computation of *D* can be illustrated by reference to the data previously reported in Table 19. First, the numbers of persons passing each item in the U and L groups are changed to percentages. Because the number of cases in each group is 20, we could divide each number by 20 and multiply the result by 100. It is easier, however, to divide 100 by 20, which gives 5, and then multiply each number by that constant. Thus, for Item 1, 15 × 5 = 75 (U group) and 7 × 5 = 35 (L group). For this item, then, *D*, is: 75 − 35 = 40. The remaining values for the seven items are given in Table 21.[3]

D can have any value between +100 and −100. If all members of the U group and none of the L group pass an item, *D* equals 100. Conversely, if all members of the L group and none of the U group pass it, *D* equals −100. If the percentages of both groups passing an item are equal, *D* will be zero. *D* has several interesting properties. It has been demonstrated (Ebel, 1965; Findley, 1956) that *D* is directly proportional to the difference between the numbers of correct and incorrect discriminations made by an item. Correct discriminations are based on the number of passes in the U group versus the number of failures in the L group; incorrect discriminations are based on the number of failures in the U group versus the number of passes in the L group. Ebel (1967) has also shown that there is a close relation between the mean *D* index of the items and the reliability coefficient of the test. The higher the mean *D*, the higher the reliability.

TABLE 21
Computation of Index of Discrimination (Data from Table 19)

Item	Percentage Passing		Difference (Index of Discrimination)
	Upper Group	Lower Group	
1	75	35	40
2	100	80	20
3	95	45	50
4	50	80	−30
5	55	55	0
6	80	45	35
7	25	0	25

Another noteworthy characteristic of *D* is one it shares with several other indices of item validity. The values of *D* are not independent of item difficulty but are biased in favor of intermediate difficulty levels. Table 22 shows the maximum possible value of *D* for items with different percentages of correct responses. If either 100 percent or 0 percent of the cases in the total sample pass an item, there can be no difference in percentage passing in U and L groups; hence *D* is zero. At the other extreme, if 50 percent pass an item, it would be possible for all the U cases and none of the L cases to pass it, thus yielding a *D* of 100 (100 − 0 = 100). If 70 percent pass, the maximum value that *D* could take can be illustrated as follows: (U) 50/50 = 100%; (L) 20/50 = 40%; *D* = 100 − 40 = 60. It will be recalled that, for most testing purposes, items closer to the 50 percent difficulty level are preferable. Hence, item validity indices that favor this difficulty level are often appropriate for item selection.

---

[3] The alert reader may have noticed that the same result can be obtained by simply multiplying the differences in the last column of Table 19 by the constant, 5.

TABLE 22
Relation of Maximum Value of *D* to Item Difficulty

Percentage Passing Item	Maximum Value of *D*
100	0
90	20
70	60
50	100
30	60
10	20
0	0

PHI COEFFICIENT. Many indices of item validity report the relationship between item and criterion in the form of a correlation coefficient. One of these is the phi coefficient (ϕ). Computed from a fourfold table, ϕ is based on the proportions of cases passing and failing an item in U and L criterion groups. Like all correlation coefficients, it yields values between +1.00 and −1.00. The ϕ coefficient assumes a genuine dichotomy in both item response and criterion variable. Consequently, it is strictly applicable only to the dichotomous conditions under which it was obtained and cannot be generalized to any underlying relationship between the traits measured by item and criterion. Like the *D* index, ϕ is biased toward the middle difficulty levels—that is, it yields the highest possible correlations for dichotomies closest to a 50–50 split.

Several computational aids are available for finding ϕ coefficients. When the number of cases in U and L criterion groups is equal, ϕ can be found with the Jurgensen tables (1947) by simply entering the percentages passing the item in U and L groups. Since in conducting an item analysis it is usually feasible to select U and L groups of equal size, the Jurgensen tables are widely used for this purpose. When the two criterion groups are unequal, ϕ can be found with another set of tables, prepared by Edgerton (1960), although their application is slightly more time consuming.

The significance level of a ϕ coefficient can be readily computed through the relation of ϕ to both Chi Square and the Normal Curve Ratio. Applying the latter, we can identify the minimum value of ϕ that would reach statistical significance at the .05 or .01 levels with the following formulas:

$$\phi_{.05} = \frac{1.96}{\sqrt{N}}$$

$$\phi_{.01} = \frac{2.58}{\sqrt{N}}$$

In these formulas, N represents the total number of cases in both criterion groups combined. Thus, if there were 50 cases in U and 50 in L groups, N would be 100 and the minimum ϕ significant at the .05 level would be $1.96 \div \sqrt{100} = .196$. Any item whose ϕ reached or exceeded .196 would thus be valid at the .05 level of significance.

BISERIAL CORRELATION. As a final example of a commonly used measure of item validity, we may consider the biserial correlation coefficient (r_{bis}), which contrasts with ϕ in two major respects. First, r_{bis} assumes a continuous and normal distribution of the trait underlying the dichotomous item response. Second, it yields a measure of item-criterion relationship that is independent of item difficulty. To compute r_{bis} directly from the data, we would need the mean criterion score of those who pass and those who fail the item, as well as the proportion of cases passing and failing the item in the entire sample and the standard deviation of the criterion scores.

Computing all the needed terms and applying the r_{bis} formula for each item can be quite time consuming. Tables have been prepared from which r_{bis} can be estimated by merely entering the percentages passing the item in the upper and lower 27 percent of the criterion group (Fan, 1952; 1954). These are the previously mentioned tables obtainable from Educational Testing Service. With these tables it is possible by entering the percentage passing in U and L groups to find three values: an estimate of p, the percentage who pass the item in the entire sample; the previously described Δ, a measure of item difficulty on an interval scale; and r_{bis} between item and criterion. These tables are only applicable when exactly 27 percent of the cases are placed in U and L groups.

There is no way of computing exact significance levels for these estimated biserial correlations, but it has been shown that their standard errors are somewhat larger than those of biserial correlations computed from all the data in the usual way. That is, the r_{bis} estimated from the Fan tables fluctuates more from sample to sample than does the r_{bis} computed by formula. With this information, one could use the standard error of r_{bis} to estimate approximately how large the correlation should be for statistical significance.[4] It should be reiterated that, with computer facilities, biserial correlations can be readily found from the responses of the total sample; and this is the preferred procedure.

INTERNAL CONSISTENCY

Item analysis is frequently conducted against total score on the test itself. This is a common practice in the case of achievement tests and especially teacher-made classroom tests, for which an external criterion is rarely available. As was noted in Chapter 6, this procedure yields a measure of internal consistency, not external validity. Such a procedure is appropriate as a refinement of content validation and of certain aspects of construct validation.

For tests requiring criterion-related validity, however, the use of total score in item analysis needs careful scrutiny. Under certain conditions, the two approaches may lead to opposite results, the items chosen on the basis of external validity being the very ones rejected on the basis of internal consistency. Let us suppose that the preliminary form of a scholastic aptitude test consists of 100 arithmetic items and 50 vocabulary items. In order to select items from this initial pool by the method of internal consistency, the biserial correlation between performance on each item and total score on the 150 items may be used.[5] It is apparent that such biserial correlations would tend to be higher for the arithmetic than for the vocabulary items, since the total score is based on twice as many arithmetic items. If it is desired to retain the 75 "best" items in the final form of the test, it is likely that most of these items will prove to be arithmetic problems. In terms of the criterion of scholastic achievement, however, the vocabulary items might have been more valid predictors than the arithmetic items. If such is the case, the item analysis will have served to lower rather than raise the validity of the test.

The practice of rejecting items that have low correlations with total score provides a means of purifying or homogenizing the test. By such a

[4] The formula for $\sigma_{r_{bis}}$ can be found in any standard statistics text, such as Guilford and Fruchter (1973, pp. 293–296).

[5] Such part-whole correlations will be somewhat inflated by the common specific and error variance in the item and the test of which it is a part. Formulas have been developed to correct for this effect (Guilford & Fruchter, 1973, pp. 454–456).

procedure, the items with the highest average intercorrelations will be retained. This method of selecting items will increase test validity only when the original pool of items measures a single trait and when this trait is present in the criterion. Some types of tests, however, measure a combination of traits required by a complex criterion. Purifying the test in such a case may reduce its criterion coverage and thus lower validity.

The selection of items to maximize test validity may be likened to the selection of tests that will yield the highest validity for a battery. It will be recalled (Ch. 7) that the test contributing most toward battery validity is one having the highest correlation with the criterion and the lowest correlation with the other tests in the battery. If this principle is applied to the selection of items, it means that the most satisfactory items are those with the highest item validities and the lowest coefficients of internal consistency. On this basis, it is possible to determine the *net effectiveness* of an item—that is, the net increase in test validity that accrues from the addition of that particular item. Thus, an item that has a high correlation with the external criterion but a relatively low correlation with total score would be preferred to one correlating highly with both criterion and test score, since the first item presumably measures an aspect of the criterion not adequately covered by the rest of the test.

It might seem that items could be selected by the same methods used in choosing tests for inclusion in a battery. Thus, each item could be correlated with the criterion and with every other item. The best items chosen by this method could then be weighted by means of a regression equation. Such a procedure, however, is neither feasible nor theoretically defensible. Not only would the computation labor be prohibitive, but interitem correlations are also subject to excessive sampling fluctuation and the resulting regression weights would be too unstable to provide a basis for item selection, unless extremely large samples are used. For these reasons, several approximation procedures have been developed for selecting items in terms of their net contribution to test validity. Some of these methods involve an empirical build-up process, whereby items are added to an ever-growing pool, and the validity of each successive composite is recomputed. Others begin with the complete set of items and reduce the pool by successive elimination of the poorest items until the desired test validity is attained. Because even these techniques require considerable computational labor, their use is practicable only when computer facilities are available (Fossum, 1973; Henrysson, 1971).

It should be noted that all techniques for selecting items in terms of their net effectiveness represent the opposite approach from that followed when items are chosen on the basis of internal consistency. In the former procedure, a high item-test correlation increases the probability that the item will be rejected; in the latter, a high item-test correlation increases the probability of its acceptance. The objectives of the two procedures are, of course, unlike. One aims to increase the breadth of criterion coverage and reduce duplication; the other attempts to raise the homogeneity of the test. Both are desirable objectives of test construction. The appropriate procedure depends largely on the nature and purpose of the test. One extreme can be illustrated by a biographical inventory, in which items can only be evaluated and selected in terms of an external criterion and the coverage of content is highly heterogeneous. The opposite extreme can be illustrated by a spelling test, whose content is highly homogeneous and in which internal consistency would be a desirable goal for item selection.

For many testing purposes, a satisfactory compromise is to sort the relatively homogeneous items into separate tests, or subtests, each of which will cover a different aspect of the criterion. Thus, breadth of coverage is achieved through a variety of tests, each yielding a relatively unambiguous score, rather than through heterogeneity of items within a single test. By such a procedure, items with low indices of internal consistency would not be discarded, but would be segregated. Within each subtest or item group, fairly high internal consistency could thus be attained. At the same time, internal consistency would not be accepted as a substitute for criterion-related validity, and some attention would be given to adequacy of coverage and to the avoidance of excessive concentration of items in certain areas.

ITEM ANALYSIS OF SPEEDED TESTS

Whether or not speed is relevant to the function being measured, item indices computed from a speeded test may be misleading. Except for items that all or nearly all subjects have had time to attempt, the item indices found from a speed test will reflect the *position* of the item in the test rather than its intrinsic difficulty or validity. Items that appear late in the test will be passed by a relatively small percentage of the total sample, because only a few persons have time to reach these items. Regardless of how easy the item may be, if it occurs late in a speeded test, it will appear difficult. Even if the item merely asked for the subject's name, the percentage of persons who pass it might be very low when the item is placed toward the end of a speeded test.

Similarly, item validities tend to be overestimated for those items that have not been reached by all subjects. Because the more proficient individuals tend to work faster, they are more likely to reach one of the later items in a speed test (Mollenkopf, 1950a). Thus, regardless of the nature of the item itself, some correlation between the item and the criterion would be obtained if the item occurred late in a speed test.

To avoid some of these difficulties, we could limit the analysis of each item to those persons who have reached the item. This is not a completely satisfactory solution, however, unless the number of persons failing to reach the item is small. Such a procedure would involve the use of a rapidly shrinking number of cases, and would thus render the results on the later items quite unreliable. Moreover, the persons on whom the later items are analyzed would probably constitute a selected sample, and hence would not be comparable to the larger samples used for the earlier items. As has already been pointed out, the faster subjects tend also to be the more proficient. The later items would thus be analyzed on a superior sample of individuals. One effect of such a selective factor would be to lower the apparent difficulty level of the later items, since the percentage passing would be greater in the selected superior group than in the entire sample. It will be noted that this is the opposite error from that introduced when the percentage passing is computed in terms of the entire sample. In that case, the apparent difficulty of items is spuriously raised.

The effect of the above procedure on indices of item validity is less obvious, but nonetheless real. It has been observed, for example, that some low-scoring examinees tend to hurry through the test, marking items almost at random in their effort to try all items within the time allowed. This tendency is much less common among high-scoring examinees. As a result, the sample on which a late-appearing item is analyzed is likely to consist of some very poor respondents, who will perform no better than chance on the item, and a larger number of very proficient and fast respondents, who are likely to answer the item correctly. In such a group, the item-criterion correlation will probably be higher than it would be in a more representative sample. In the absence of such random respondents, on the other hand, the sample on which the later items are analyzed will cover a relatively narrow range of ability. Under these conditions, the validities of the later items will tend to be lower than they would be if computed on the entire unselected sample.

The anticipated effects of speed on indices of item difficulty and item validity have been empirically verified, both when item statistics are computed with the entire sample (Wesman, 1949) and when they are computed with only those persons who attempt the item (Mollenkopf, 1950a). In the latter study, comparable groups of high school students were given two forms of a verbal test and two forms of a mathematics test. Each of the two forms contained the same items as the other, but items occurring early in one form were placed late in the other. Each form was administered with a short time limit (speed conditions) and with a very liberal time limit (power conditions). Various intercomparisons were thus possible between forms and timing conditions. The results clearly showed that the position of an item in the speed tests affected its indices of difficulty and validity. When the same item occurred later in a speeded test, it was passed by a greater percentage of those attempting it, and it yielded a higher item-criterion correlation.

The difficulties encountered in the item analysis of speeded tests are fundamentally similar to those discussed in Chapter 5 in connection with the reliability of speeded tests. Various solutions, both empirical and statistical, have been developed for meeting these difficulties. One empirical solution is to administer the test with a longer time limit to the group on which item analysis is to be carried out. This solution is satisfactory provided that speed itself is not an important aspect of the ability to be measured by the test. Apart from the technical problems presented by specific tests, however, it is well to keep in mind that item-analysis data obtained with speeded tests are suspect and call for careful scrutiny.

CROSS VALIDATION

MEANING OF CROSS VALIDATION. It is essential that test validity be computed on a different sample of persons from that on which the items were selected. This independent determination of the validity of the entire test is known as cross validation (Mosier, 1951). Any validity coefficient computed on the same sample that was used for item-selection purposes will capitalize on chance errors within that particular sample and will consequently be spuriously high. In fact, a high validity coefficient could result under such circumstances even when the test has no validity at all in predicting the particular criterion.

Let us suppose that out of a sample of 100 medical students, the 30 with the highest and the 30 with the lowest medical school grades have been chosen to represent contrasted criterion groups. If, now, these two groups are compared in a number of traits actually irrelevant to success in medical school, certain chance differences will undoubtedly be found. Thus, there might be an excess of private-school graduates and of red-haired persons within the upper criterion group. If we were to assign each individual a score by crediting him with one point for private-school graduation and one point for red hair, the mean of such scores would undoubtedly be higher in the upper than in the lower criterion group. This is not evidence for the validity of the predictors, however, since such a validation process is based on a circular argument. The two predictors were chosen in the first place on the basis of the chance variations that characterized this particular sample. And the *same* chance differences are operating to produce the mean differences in total score.

When tested in another sample, however, the chance differences in frequency of private-school graduation and red hair are likely to disappear or be reversed. Consequently, the validity of the scores will collapse.

AN EMPIRICAL EXAMPLE. A specific illustration of the need for cross validation is provided by an early investigation conducted with the Rorschach inkblot test (Kurtz, 1948). In an attempt to determine whether the Rorschach could be of any help in selecting sales managers for life insurance agencies, this test was administered to 80 such managers. These managers had been carefully chosen from several hundred employed by eight life insurance companies, so as to represent an upper criterion group of 42 considered very satisfactory by their respective companies, and a lower criterion group of 38 considered unsatisfactory. The 80 test records were studied by a Rorschach expert, who selected a set of 32 signs, or response characteristics, occurring more frequently in one criterion group than in the other. Signs found more often in the upper criterion group were scored +1 if present and 0 if absent; those more common in the lower group were scored −1 or 0. Since there were 16 signs of each type, total scores could range theoretically from −16 to +16.

When the scoring key based on these 32 signs was reapplied to the original group of 80 persons, 79 of the 80 were correctly classified as being in the upper or lower group. The correlation between test score and criterion would thus have been close to 1.00. However, when the test was cross-validated on a second comparable sample of 41 managers, 21 in the upper and 20 in the lower group, the validity coefficient dropped to a negligible .02. It was thus apparent that the key developed in the first sample had no validity for selecting such personnel.

AN EXAMPLE WITH CHANCE DATA. That the use of a single sample for item selection and test validation can produce a completely spurious validity coefficient under pure chance conditions was vividly demonstrated by Cureton (1950). The criterion to be predicted was the grade-point average of each of 29 students registered in a psychology course. This criterion was dichotomized into grades of B or better and grades below B. The "items" consisted of 85 tags, numbered from 1 to 85 on one side. To obtain a test score for each subject, the 85 tags were shaken in a container and dropped on the table. All tags that fell with numbered side up were recorded as indicating the presence of that particular item in the student's test performance. Twenty-nine throws of the 85 tags thus provided complete records for each student, showing the presence or absence of each item or response sign. Because of the procedure followed in generating these chance scores, Cureton facetiously named the test the "B-Projective Psychokinesis Test."

An item analysis was then conducted, with each student's grade-point average as the criterion. On this basis, 24 "items" were selected out of the 85. Of these, 9 occurred more frequently among the students with an average grade of B or better and received a weight of +1; 15 occurred more frequently among the students with an average grade below B and received a weight of −1. The sum of these item weights constituted the total score for each student. Despite the known chance derivation of these "test scores," their correlation with the grade criterion in the original group of 29 students proved to be .82. Such a finding is similar to that obtained with the Rorschach scores in the previously cited study. In both instances, the apparent correspondence between test score and criterion resulted from the utilization of the same chance differences both in selecting items and in determining validity of total test scores.

CONDITIONS AFFECTING VALIDITY SHRINKAGE. The amount of shrinkage of a validity coefficient in cross validation depends in part on the size of the original item pool and the proportion of items retained. When the number of original items is large and the proportion retained is small, there is more opportunity to capitalize on chance differences and thus obtain a spuriously high validity coefficient. Another condition affecting amount of shrinkage in cross validation is size of sample. Since spuriously high validity in the initial sample results from an accumulation of sampling errors, smaller groups (which yield larger sampling errors) will exhibit greater validity shrinkage.

If items are chosen on the basis of previously formulated hypotheses, derived from psychological theory or from past experience with the criterion, validity shrinkage in cross validation will be minimized. For example, if a particular hypothesis required that the answer "Yes" be more frequent among successful students, then the item would *not* be retained if a significantly larger number of "Yes" answers were given by the *unsuccessful* students. The opposite, blindly empirical approach would be illustrated by assembling a miscellaneous set of questions with little regard to their relevance to the criterion behavior, and then retaining all items yielding significant positive or negative correlations with the criterion. Under the latter circumstances, we would expect much more shrinkage than under the former. In summary, shrinkage of test validity in cross validation will be greatest when samples are small, the initial item pool is large, the proportion of items retained is small, and items are assembled without previously formulated rationale.

ITEM-BY-GROUP INTERACTION

EXPLORATORY STUDIES. Insofar as diverse cultures or subcultures foster the development of different skills and knowledge, these differences will be reflected in test scores. An individual's general level of performance will be higher in those aptitudes stimulated and encouraged by his particular experiential background. A further question pertains to the *relative difficulty* of items for groups with dissimilar cultural backgrounds. If difficulty is measured in the usual way, in terms of percentage of respondents passing each item, will the rank order of the items be the same across groups, regardless of overall level of performance? Early investigations of this question with urban and rural children revealed a number of significant differences in relative item difficulties on the Stanford-Binet (Jones, Conrad, & Blanchard, 1930) and on a general information test (Shimberg, 1929).

A comprehensive test of group differences in relative item difficulties is provided by measures of item × group interaction, computed through an analysis of variance. Another procedure is to correlate either the per-

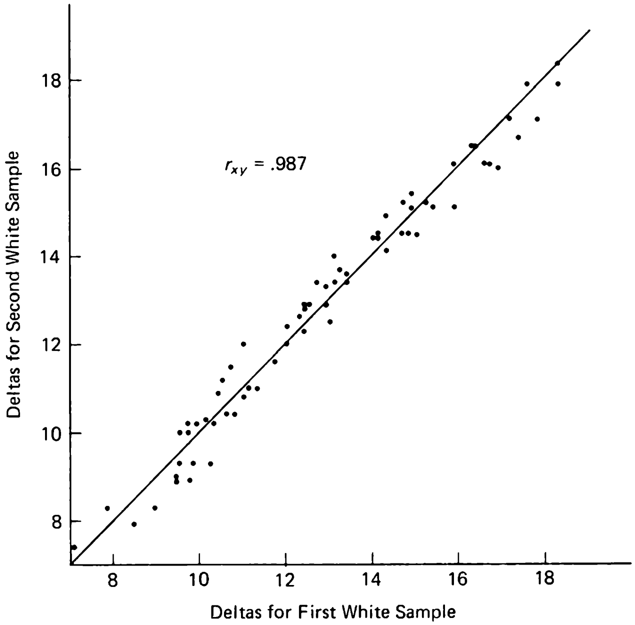


FIG. 27. Bivariate Distribution of Item Difficulties of Preliminary Scholastic Aptitude Test for Two Random Samples of White High School Students.

(From Angoff & Ford, 1973, p. 99. Reproduced by permission of National Council on Measurement in Education.)

centage passing or the delta values (Δ) of the same items in two groups. If there is no significant item × group interaction, i.e., if the relative difficulties of items are the same for both groups, this correlation should be close to 1.00. These more sophisticated statistical techniques have been employed in studies with the College Board's Preliminary Scholastic Aptitude Test, administered to high school students. Relative item difficulties were investigated with reference to ethnic, socioeconomic, and urban-rural categories (Angoff & Ford, 1973; Cleary & Hilton, 1968).

The results show some significant though small item × group interactions. Correlations between the delta values for two ethnic groups were slightly lower than the correlations for two random samples of the same ethnic group. Two of the bivariate distributions of such correlations are illustrated in Figures 27 and 28. When two random samples of white high school students were compared (Fig. 27), the item deltas were closely similar, yielding a correlation of .987. When black and white samples were compared (Fig. 28), the items not only proved more difficult as a whole for the black students, but they also showed more discrepancies in relative difficulty, as indicated by a correlation of .929.

Efforts to identify reasons for these differences led to two suggestive

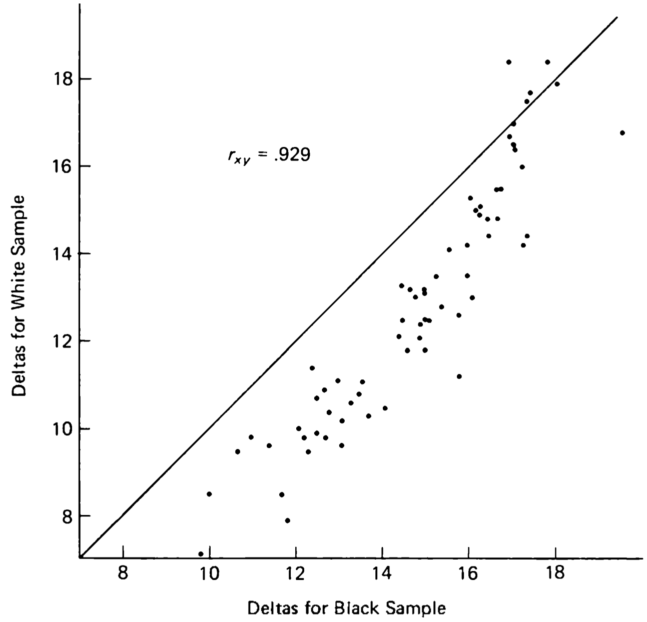


FIG. 28. Bivariate Distribution of Item Difficulties of Preliminary Scholastic Aptitude Test for Random Samples of Black and White High School Students. (From Angoff & Ford, 1973, p. 100. Reproduced by permission of National Council on Measurement in Education.)

findings. First, examination of item content failed to reveal any relation between the affected items and known differences in the experiential backgrounds of the groups. Second, equating the groups on a related cognitive variable reduced both the group difference in mean scores and the item × group interaction. The latter finding suggests that the relative difficulty of items depends at least in part on the absolute performance level in the ability measured by the test. It is possible, for example, that persons at different aptitude levels utilize different work methods, problem-solving techniques, or cognitive skills in responding to the same items. Those items that prove relatively difficult when solved by method A could prove relatively easy when solved by method B, and vice versa.

It should be added that all the techniques used in studying item × group interactions in ability tests are also applicable to personality tests. In the latter tests, what is measured is not the difficulty of items but the relative frequency in choice of specific response options, as on an attitude scale or personality inventory.

ITEM SELECTION TO MINIMIZE OR MAXIMIZE GROUP DIFFERENCES. In the construction of certain tests, item × group interactions have been used as one basis for the selection of items. In the development of the Stanford-Binet, for example, an effort was made to exclude any item that favored either sex significantly, on the assumption that such items might reflect purely fortuitous and irrelevant differences in the experiences of the two sexes (McNemar, 1942, Ch. 5). Owing to the limited number of items available for each age level, however, it was not possible to eliminate all sex-differentiating items. In order to rule out sex differences in total score, therefore, the remaining sex-differentiating items were balanced, approximately the same number favoring boys and girls.

No generalization can be made regarding the elimination of sex differences, or any other group differences, in the selection of test items. While certain tests, like the Stanford-Binet, have sought to equalize the performance of the two sexes, others have retained such differences and report separate norms for the two sexes. This practice is relatively common in the case of special aptitude tests, in which fairly large differences in favor of one or the other sex have been consistently found.

Under certain circumstances, moreover, items may be chosen, not to minimize, but to maximize, sex differentiation. An example of the latter procedure is to be found in the masculinity-femininity scales developed for use with several personality inventories (to be discussed in Ch. 17). Since the purpose of these scales is to measure the degree to which an individual's responses agree with those characteristic of men or of women in our culture, only those items that differentiate significantly between the sexes are retained.

A similar diversity of procedure can be found with reference to other group differences in item performance. In the development of a socioeconomic status scale for the Minnesota Multiphasic Personality Inventory, only those items were retained that differentiated significantly between the responses of high school students in two contrasted socioeconomic groups (Gough, 1948). Cross validation of this status scale on a new sample of high school students yielded a correlation of .50 with objective indices of socioeconomic status. The object of this test is to determine the degree to which an individual's emotional and social responses resemble those characteristic of persons in upper or lower socioeconomic levels, respectively. Hence, those items showing the maximum differentiation between social classes were included in the scale, and those showing little or no differentiation were discarded. This procedure is similar to that followed in the development of masculinity-femininity scales. It is apparent that in both types of tests the group differentiation constitutes the criterion in terms of which the test is validated. In such cases, socioeconomic level and sex, respectively, represent the most relevant variables on the basis of which items can be chosen.

Examples of the opposite approach to socioeconomic or cultural differentials in test responses can also be found. An extensive project on such cultural differentials in intelligence test items was conducted at the University of Chicago (Eells et al., 1951). These investigators believed that most intelligence tests might be unfair to children from lower socioeconomic levels, since many of the test items presuppose information, skills, or interests typical of middle-class children. To obtain evidence for such a hypothesis, a detailed item analysis was conducted on eight widely used group intelligence tests. For each item, the frequencies of correct responses by children in higher and lower socioeconomic levels were compared. Following this investigation, two members of the research team prepared a special test designed to be "fair" to lower-class urban American children. In the construction of this test, an effort was made to exclude the types of items previously found to favor middle-class children.[6]

As in the case of sex differences, no rigid policy can be laid down regarding items that exhibit cultural differentiation. Certain basic facts of test construction and interpretation should, however, be noted. First, whether items that differentiate significantly between certain groups are retained or discarded should depend on the purpose for which the test is designed. If the criteria to be predicted show significant differences between the sexes, socioeconomic groups, or other categories of persons, then it is to be expected that the test items will also exhibit such group differences. To eliminate items showing these differences might serve only to lower the validity of the test for predicting the given criteria (see Anastasi, 1966). In the second place, tests designed to measure an individual's resemblance to one or another group should obviously magnify the differentiation between such groups. For these tests, items showing the largest group differences in response should be chosen, as in the case of the masculinity-femininity and social status scales cited above.

---

[6] Known as the Davis-Eells Games, this test was subsequently discontinued because it proved unsatisfactory in a number of ways, including low validity in predicting academic achievement and other practical criteria. Moreover, the anticipated advantage of lower-class children on this test did not hold up in other samples.

The third point is of primary concern, not to the test constructor, but to the test user and the general student of psychology who wishes to interpret test results properly. Tests whose items have been selected with reference to the responses of any special groups cannot be used to compare such groups. For example, the statement that boys and girls do not differ significantly in Stanford-Binet IQ provides no information whatever regarding sex differences. Since sex differences were deliberately eliminated in the process of selecting items for the test, their absence from the final scores merely indicates that this aspect of test construction was successfully executed. Similarly, lack of socioeconomic differences on a test constructed so as to eliminate such differences would provide no information on the relative performance of groups varying in socioeconomic status.

Tests designed to maximize group differentiation, such as the masculinity-femininity and social status scales, are equally unsuitable for group comparisons. In these cases, the sex or socioeconomic differentiation in personality characteristics would be artificially magnified. To obtain an unbiased estimate of the existing group differences, the test items must be selected without reference to the responses of such groups. The principal conclusion to be drawn from the present discussion is that proper interpretation of scores on any test requires a knowledge of the basis on which items were selected for that test.

PART 3

Tests of General Intellectual Level

CHAPTER 9

Individual Tests

IN PART 2, we were concerned with the major principles of psychological testing. We are now ready to apply these principles to the evaluation of specific tests. We now know what questions to ask about each test and where to look for the answers. The test manuals, the *Mental Measurements Yearbooks,* appropriate journals, and other sources described in Chapter 1 may be consulted to obtain information regarding any of the tests cited.

The purpose of the remaining parts of the book is twofold. One objective is to afford an opportunity to observe the application of testing principles to a wide variety of tests. Another is to acquaint the reader with a few outstanding tests in each of the major areas. No attempt will be made to provide a comprehensive survey of available tests within any area. Such a survey would be outside the scope of this book. Moreover, it would probably be outdated before publication, because of the rapidity with which new tests appear. For these reasons, the discussion will concentrate on a few representative tests in each category, chosen either because of their widespread use or because they illustrate important developments in testing procedure. We shall consider tests of general intellectual level in Part 3, tests of separate abilities in Part 4, and personality tests in Part 5.

A classified list of representative tests in all areas covered in this text will be found in Appendix E. This list includes not only all currently available tests cited in the text, but also some additional examples of well-known tests in each category. With each test, the table gives the name of the publisher and the entry number in the latest *Mental Measurements Yearbook* where information about the test can be located. Unless otherwise indicated, all data about tests discussed in the text are taken from the test manual or technical supplements supplied by the test publishers. As a further aid to the reader's own evaluation of tests, Appendix C contains a suggested outline for test evaluation. The APA *Standards* (1974) provide more detailed guidelines for this purpose.

Traditionally called "intelligence tests," the types of tests to be discussed in Part 3 are the direct descendants of the original Binet scales. Such tests are designed for use in a wide variety of situations and are validated against relatively broad criteria. They characteristically provide a single score, such as an IQ, indicating the individual's general intellectual level. A typical approach is to arrive at this global estimate of intellectual performance by "the sinking of shafts at critical points" (Terman & Merrill, 1937, p. 4). In other words, a wide variety of tasks is presented to the subject in the expectation that an adequate sampling of all important intellectual functions will thus be covered. In actual practice, the tests are usually overloaded with certain functions, such as verbal ability, and completely omit others.

Because so many intelligence tests are validated against measures of academic achievement, they are often designated as tests of scholastic aptitude. Intelligence tests are frequently employed as preliminary screening instruments, to be followed by tests of special aptitudes. This practice is especially prevalent in the testing of normal adolescents or adults for educational and vocational counseling, personnel selection, and similar purposes. Another common use of general intelligence tests is to be found in clinical testing, especially in the identification and classification of the mentally retarded. For clinical purposes, individual tests are generally employed. Among the most widely used individual intelligence tests are the Stanford-Binet and Wechsler scales discussed in this chapter.

STANFORD-BINET INTELLIGENCE SCALE

EVOLUTION OF THE SCALES. The original Binet-Simon scales have already been described briefly in Chapter 1. It will be recalled that the 1905 scale consisted simply of 30 short tests, arranged in ascending order of difficulty. The 1908 scale was the first age scale; and the 1911 scale introduced minor improvements and additions. The age range covered by the 1911 revision extended from 3 years to the adult level. Among the many translations and adaptations of the early Binet tests were a number of American revisions, of which the most viable has been the Stanford-Binet.[1] The first Stanford revision of the Binet-Simon scales, prepared by Terman and his associates at Stanford University, was published in 1916 (Terman, 1916). This revision introduced so many changes and additions as to represent virtually a new test. Over one third of the items were new, and a number of old items were revised, reallocated to different age levels, or discarded. The entire scale was restandardized on an American sample of approximately one thousand children and four hundred adults.

---

[1] A detailed account of the Binet-Simon scales and of the development, use, and clinical interpretation of the Stanford-Binet can be found in Sattler (1974, Chs. 8–10).

Detailed instructions for administering and scoring each test[2] were provided, and the IQ was employed for the first time in any psychological test.

The second Stanford revision, appearing in 1937, consisted of two equivalent forms, L and M (Terman & Merrill, 1937). In this revision, the scale was greatly expanded and completely restandardized on a new and carefully chosen sample of the U.S. population. The 3,184 subjects employed for this purpose included approximately 100 children at each half-year interval from 1½ to 5½ years, 200 at each age from 6 to 14, and 100 at each age from 15 to 18. All subjects were within one month of a birthday (or half-year birthday) at the time of testing, and every age group contained an equal number of boys and girls. From age 6 up, most subjects were tested in school, although a few of the older subjects were obtained outside of school in order to round out the sampling. Preschool children were contacted in a variety of ways, many of them being siblings of the schoolchildren included in the sample. Despite serious efforts to obtain a representative cross-section of the population, the sampling was somewhat higher than the U.S. population in socioeconomic level, contained an excess of urban cases, and included only native-born whites.

A third revision, published in 1960, provided a single form (L-M) incorporating the best items from the two 1937 forms (Terman & Merrill, 1960). Without introducing any new content, it was thus possible to eliminate obsolescent items and to relocate items whose difficulty level had altered during the intervening years owing to cultural changes. In preparing the 1960 Stanford-Binet, the authors were faced with a common dilemma of psychological testing. On the one hand, frequent revisions are desirable in order to profit from technical advances and refinements in test construction and from prior experience in the use of the test, as well as to keep test content up to date. The last-named consideration is especially important for information items and for pictorial material which may be affected by changing fashions in dress, household appliances, cars, and other common articles. The use of obsolete test content may seriously undermine rapport and may alter the difficulty level of items. On the other hand, revision may render much of the accumulated data inapplicable to the new form. Tests that have been widely used for many years have acquired a rich body of interpretive material which should be carefully weighed against the need for revision. It was for these reasons that the authors of the Stanford-Binet chose to condense the two earlier forms into one, thereby steering a course between the twin hazards of obsolescence and discontinuity. The loss of a parallel form was not too great a price to pay for accomplishing this purpose. By 1960 there was less need for an alternate form than there had been in 1937 when no other well-constructed individual intelligence scale was available.

---

[2] The items in the Binet scales are commonly called "tests," since each is separately administered and may contain several parts.

In the preparation of the 1960 Stanford-Binet, items were selected from forms L and M on the basis of the performance of 4,498 subjects, aged 2½ to 18 years, who had taken either or both forms of the test between 1950 and 1954. The subjects were examined in six states situated in the Northeast, in the Midwest, and on the West Coast. Although these cases did not constitute a representative sampling of American schoolchildren, care was taken to avoid the operation of major selective factors.[3] The 1960 Stanford-Binet did *not* involve a restandardization of the normative scale. The new samples were utilized only to identify changes in item difficulty over the intervening period. Accordingly, the difficulty of each item was redetermined by finding the percentage of children passing it at successive *mental ages* on the 1937 forms. Thus for purposes of item analysis the children were grouped, not according to their chronological age, but according to the mental age they had obtained on the 1937 forms. Consequently, mental ages and IQ's on the 1960 Form L-M were still expressed in terms of the 1937 normative sample.

The next stage was the 1972 restandardization of Form L-M (Terman & Merrill, 1973, Part 4). This time the test content remained unchanged,[4] but the norms were derived from a new sample of approximately 2,100 cases tested during the 1971–1972 academic year. To achieve national representativeness despite the practical impossibility of administering individual tests to very large samples, the test publishers took advantage of a sample of approximately 20,000 children at each age level, employed in the standardization of a group test (Cognitive Abilities Test). This sample of some 200,000 schoolchildren in grades 3 through 12 was chosen from communities stratified in terms of size, geographical region, and economic status, and included black, Mexican-American, and Puerto Rican children.

The children to be tested with the Stanford-Binet were identified through their scores on the verbal battery of the Cognitive Abilities Test, so that the distribution of scores in this subsample corresponded to the national distribution of the entire sample. The only cases excluded were those for whom the primary language spoken in the home was not English. To cover ages 2 to 8, the investigators located siblings of the group-tested children, choosing each child on the basis of the Cognitive Abilities Test score obtained by his or her older sibling. Additional cases at the upper ages were recruited in the same way. The Stanford-Binet sample included approximately 100 cases in each half-year age group from 2 to 5½ years, and 100 at each year group from 6 to 18.

---

[3] For special statistical analyses, there were two additional samples of California children, including 100 6-year-olds stratified with regard to father's occupation and 100 15-year-olds stratified with regard to both father's occupation and grade distribution.

[4] With only two very minor exceptions: the picture on the "doll card" at age II was updated; and the word "charcoal" was permitted as a substitute for "coal" in the Similarities test at age VII.

In comparison with the 1937 norms, the 1972 norms are based on a more representative sample, as well as being updated and hence reflecting any effects of intervening cultural changes on test performance. It is interesting to note that the later norms show some improvement in test performance at all ages. The improvement is substantial at the preschool ages, averaging about 10 IQ points. The test authors attribute this improvement to the impact of radio and television on young children and the increasing literacy and educational level of parents, among other cultural changes. There is a smaller but clearly discernible improvement at ages 15 and over which, as the authors suggest, may be associated with the larger proportion of students who continued their education through high school in the 1970s than in the 1930s.

ADMINISTRATION AND SCORING. The materials needed to administer the Stanford-Binet are shown in Figure 29. They include a box of standard toy objects for use at the younger age levels, two booklets of printed cards, a record booklet for recording responses, and a test manual. The tests are grouped into age levels extending from age II to superior adult. Between the ages of II and V, the test proceeds by half-year intervals. Thus, there is a level corresponding to age II, one to age II-6, one to age III, and so forth. Because progress is so rapid during these early ages, it proved feasible and desirable to measure change over six-month intervals. Between V and XIV the age levels correspond to yearly intervals. The remaining levels are designated as Average Adult and Superior Adult levels I, II, and III. Each age level contains six tests, with the exception of the Average Adult level, which contains eight.

The tests within any one age level are of approximately uniform difficulty and are arranged without regard to such residual differences in difficulty as may be present. An *alternate* test is also provided at each age level. Being of approximately equivalent difficulty, the alternate may be substituted for any of the tests in the level. Alternates are used if one of the regular tests must be omitted because special circumstances make it inappropriate for the individual or because some irregularity interfered with its standardized administration.

Four tests in each year level were selected on the basis of validity and representativeness to constitute an *abbreviated scale* for use when time does not permit the administration of the entire scale. These tests are marked with an asterisk on the record booklets. Comparisons between full-scale and abbreviated-scale IQ's on a variety of groups show a close correspondence between the two, the correlations being approximately as high as the reliability coefficient of the full scale (Himelstein, 1966; Sattler, 1974, p. 116; Terman & Merrill, 1973, pp. 61–62). The mean IQ, however, tends to run slightly lower on the short scale. This discrepancy is also found when the numbers of persons scoring higher on each version are compared. Over 50 percent of the subjects receive lower IQ's on the short version, while only 30 percent score higher.

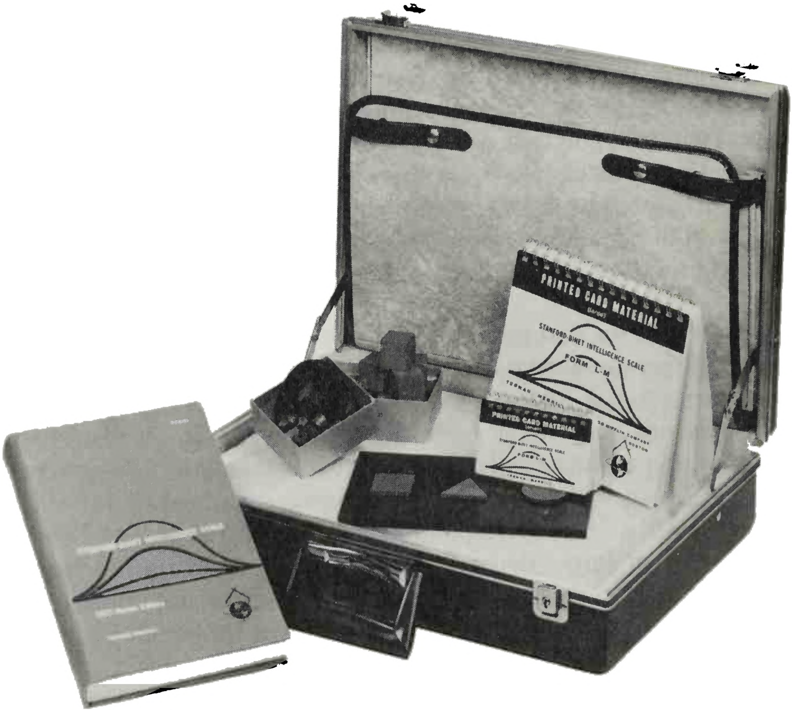


Fig. 29. Test Materials Employed in Administering the Stanford-Binet.
(Courtesy Houghton Mifflin Company.)

In common with most individual intelligence tests, the Stanford-Binet requires a highly trained examiner. Both administration and scoring are fairly complicated for many of the tests. Considerable familiarity and experience with the scale are therefore required for a smooth performance. Hesitation and fumbling may be ruinous to rapport. Slight inadvertent changes in wording may alter the difficulty of items. A further complication is presented by the fact that tests must be scored as they are administered, since the subsequent conduct of the examination depends on the child's performance on previously administered levels.

Many clinicians regard the Stanford-Binet not only as a standardized test, but also as a clinical interview. The very characteristics that make this scale so difficult to administer also create opportunities for interaction between examiner and subject, and provide other sources of clues for the experienced clinician. Even more than most other individual tests, the Stanford-Binet makes it possible to observe the subject's work methods, his approach to a problem, and other qualitative aspects of performance. The examiner may also have an opportunity to judge certain personality characteristics, such as activity level, self-confidence, persistence, and ability to concentrate. Any qualitative observations made in the course of Stanford-Binet administration should, of course, be clearly recognized as such and ought not to be interpreted in the same manner as objective test scores. The value of such qualitative observations depends to a large extent on the skill, experience, and psychological sophistication of the examiner, as well as on his awareness of the pitfalls and limitations inherent in this type of observation. The types of clinical observations that can be made during an individual intelligence examination are richly illustrated by Moriarty (1960, 1961, 1966), who sees in the testing session an opportunity to investigate the child's behavior in meeting a challenging, demanding, difficult, or frustrating situation.

In taking the Stanford-Binet, no one subject tries all items. Each individual is tested only over a range of age levels suited to his own intellectual level. Testing usually requires no more than thirty to forty minutes for younger children and not more than one hour and a half for older examinees. The standard procedure is to begin testing at a level slightly below the expected mental age of the examinee. Thus, the first tests given should be easy enough to arouse confidence, but not so easy as to cause boredom and annoyance. If the individual fails any test within the year level first administered, the next lower level is given. This procedure continues until a level is reached at which all tests are passed. This level is known as the *basal age.* Testing is then continued upward to a level at which all tests are failed, designated as the *ceiling age.* When this level is reached, the test is discontinued.

Individual Stanford-Binet items, or tests, are scored on an all-or-none basis. For each test, the minimal performance that constitutes "passing" is specified in the manual. For example, in identifying objects by use at year level II-6, the child passes if he correctly identifies three out of six designated objects; in answering comprehension questions at year level VIII, any four correct answers out of six represent a passing performance. Certain tests appear in identical form at different year levels, but are scored with a different standard of passing. Such tests are administered only once, the individual's performance determining the year level at which they are credited. The vocabulary test, for example, may be scored anywhere from level VI to Superior Adult III, depending on the number of words correctly defined.

The items passed and failed by any one individual will show a certain amount of *scatter* among adjacent year levels. We do not find that examinees pass all tests at or below their mental age level and fail all tests above such a level. Instead, the successfully passed tests are spread over several year levels, bounded by the subject's basal age at one extreme and his ceiling age at the other. The subject's mental age on the Stanford-Binet is found by crediting him with his basal age and adding to that age further months of credit for every test passed beyond the basal level. In the half-year levels between II and V, each of the six tests counts as one month; between VI and XIV, each of the six tests corresponds to two months of credit. Since each of the adult levels (AA, SA I, SA II, and SA III) covers more than one year of mental age, the months of credit for each test are adjusted accordingly. For example, the Average Adult level includes eight tests, each of which is credited with two months; the Superior Adult I level contains six tests, each receiving four months.

The highest mental age theoretically attainable on the Stanford-Binet is 22 years and 10 months. Such a score is not, of course, a true mental age, but a numerical score indicating degree of superiority above the Average Adult performance. It certainly does not correspond to the achievement of the average 22-year-old; according to the 1972 norms, the average 22-year-old obtains a mental age of 16–8. For any adult over 18 years of age, a mental age of 16–8 yields an IQ of 100 on this scale. In fact, above 13 years, mental ages cease to have the same significance as they do at lower levels, since it is just beyond 13 that the mean MA begins to lag behind CA on this scale. The Stanford-Binet is not suitable for adult testing, especially within the normal and superior range. Despite the three Superior Adult levels, there is insufficient ceiling for most superior adults or even for very superior adolescents (Kennedy et al., 1960). In such cases, it is often impossible to reach a ceiling age level at which all tests are failed. Moreover, most of the Stanford-Binet tests have more appeal for children than for adults, the content being of relatively little interest to most adults.

NORMATIVE INTERPRETATION. A major innovation introduced in the 1960 Stanford-Binet was the substitution of deviation IQ's for the ratio IQ's used in the earlier forms. These deviation IQ's are standard scores with a mean of 100 and an *SD* of 16. As explained in Chapter 4, the principal advantage of this type of IQ is that it provides comparable scores at all age levels, thus eliminating the vagaries of ratio IQ's. Despite the care with which the 1937 scales were developed in the effort to obtain constant IQ variability at all ages, the *SD*'s of ratio IQ's on these scales fluctuated from a low of 13 at age VI to a high of 21 at age II-6. Thus, an IQ of 113 at age VI corresponded to an IQ of 121 at age II-6. Special correction tables were developed to adjust for the major IQ variations in the 1937 scales (McNemar, 1942, pp. 172–174). All these difficulties were circumvented in the 1960 form through the use of deviation IQ's, which automatically have the same *SD* throughout the age range.

As an aid to the examiner, Pinneau prepared tables in which deviation IQ's can be looked up by entering MA and CA in years and months. These Pinneau tables are reproduced in the Stanford-Binet manual (Terman & Merrill, 1973). The latest manual includes both the 1972 and the 1937 normative IQ tables. For most testing purposes, the 1972 norms are appropriate, showing how the child's performance compares with that of others of his own age in his generation. To provide comparability with IQ's obtained earlier, however, the 1937 norms are more suitable. They would thus be preferred in a continuing longitudinal study, or in comparing an individual's IQ with the IQ he obtained on the Stanford-Binet at a younger age. When used in this way, the 1937 standardization sample represents a fixed reference group, just as the students taking the College Board Scholastic Aptitude Test in 1941 provide a fixed reference group for that test (see Ch. 4).

Although the deviation IQ is the most convenient index for evaluating an individual's standing in his age group, the MA itself can serve a useful function. To say that a 6-year-old child performs as well as a typical 8-year-old usually conveys more meaning to a layman than saying he has an IQ of 137. A knowledge of the child's MA level also facilitates an understanding of what can be expected of him in terms of educational achievement and other developmental norms of behavior. It should be noted, however, that the MA's obtained on the Stanford-Binet are still expressed in terms of the 1937 norms. It is only the IQ tables that incorporate the updated 1972 norms. Reference to these tables will show, for example, that if a child whose CA is 5–0 obtains an MA of 5–0, his IQ is *not* 100. To receive an IQ of 100 with the 1972 norms, this child would need an MA of 5–6.

One of the advantages of the Stanford-Binet derives from the mass of interpretive data and clinical experience that have been accumulated regarding this test. For many clinicians, educators, and others concerned with the evaluation of general ability level, the Stanford-Binet IQ has become almost synonymous with intelligence. Much has been learned about what sort of behavior can be expected from a child with an IQ of 50 or 80 or 120 on this test. The distributions of IQ's in the successive standardization samples (1916, 1937, 1972) have provided a common frame of reference for the interpretation of IQ's.

Because of the size of the error of measurement of a Stanford-Binet IQ, it is customary to allow approximately a 10-point band on either side of the obtained IQ for chance variation. Thus any IQ between 90 and 110 is considered equivalent to the average IQ of 100. IQ's above 110 represent superior deviations, those below 90 inferior deviations. There is no generally accepted frame of reference for classifying superior IQ's. It may be noteworthy, however, that in the classical, long-term investigation of gifted children by Terman and his co-workers a minimum IQ of 140 was required for inclusion in the principal part of the project (Terman & Oden, 1959).

At the other end of the scale, a widely used educational classification of mental retardates recognizes the educable, trainable, and custodial categories. The educable group, in the IQ range from 50 to 75, can advance to at least the third grade in academic work—and possibly as high as the sixth grade—if taught in a specially adapted classroom situation. The trainable group, with IQ's between 25 and 50, can be taught self-care and social adjustment in a protected environment. Those below IQ 25 generally require custodial and nursing care.

In its manual on terminology and classification, the American Association on Mental Deficiency (AAMD) lists four levels of mental retardation, defined more precisely in terms of *SD* units. This classification is given in Table 23, together with the Stanford-Binet IQ ranges corresponding to each level and the expected percentage of cases. It will be noted that the classification is based on a division of the lower portion of the normal distribution curve into steps of 1 *SD* each, beginning at −2 *SD*. The advantage of such a classification is that it can be readily translated into standard scores or deviation IQ's in any scale. Since the Stanford-Binet deviation IQ scale has an *SD* of 16, the mild level, extending from

TABLE 23
Levels of Mental Retardation as Defined in Manual of American Association on Mental Deficiency

(Data in first two columns from Grossman, 1973, p. 18. Reprinted by permission of the American Association on Mental Deficiency)

Level	Cutoff Points (in *SD* units from Mean)	Range of Stanford-Binet IQ (*SD* = 16)	Percentage of Cases
Mild	−2	68–52	2.14
Moderate	−3	51–36	0.13
Severe	−4	35–20	0.003
Profound	−5	19 and below	0.00003

−2 *SD* down to −3 *SD*, ranges from 68 (100 − 2 × 16) to 52 (100 − 3 × 16). The other IQ ranges can be found in a similar manner. The percentages of cases at each level are those expected in a normal distribution (see Fig. 6, Ch. 4). They agree quite closely with the percentages of persons at these IQ levels found empirically in the general population. The frequency of mental retardation in the general population is usually estimated as close to 2 percent. The Stanford-Binet manual contains still another classification of levels of mental retardation, based on somewhat different IQ limits, which has been widely used as an interpretive frame of reference by clinical psychologists (Terman & Merrill, 1973, p. 18).

The use of such classifications of IQ levels, although of unquestionable help in standardizing the interpretation of test performance, carries certain dangers. Like all classifications of persons, it should not be rigidly applied, nor used to the exclusion of other data about the individual. There are, of course, no sharp dividing lines between the "mentally retarded" and the "normal" or between the "normal" and the "superior." Individuals with IQ's of 60 have been known to make satisfactory adjustments to the demands of daily living, while some with IQ's close to 100 may require institutional care.

Decisions regarding institutionalization, parole, discharge, or special training of mental retardates must take into account not only IQ but also social maturity, emotional adjustment, physical condition, and other circumstances of the individual case. The AAMD defines mental retardation as "significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior, and manifested during the developmental period" (Grossman, 1973, p. 11). This definition is further explicated in the stipulation that a child should not be classified as mentally retarded unless he is deficient in *both* intellectual functioning, as indicated by IQ level, and in adaptive behavior, as measured by such instruments as the Vineland Social Maturity Scale or the AAMD Adaptive Behavior Scales (to be discussed in Chapter 10).

Nor is high IQ synonymous with genius. Persons with IQ's of 160 do occasionally lead undistinguished lives, while some with IQ's much closer to 100 may make outstanding contributions. High-level achievement in specific fields may require special talents, originality, persistence, singleness of purpose, and other propitious emotional and motivational conditions.

RELIABILITY. The reliability of the 1937 Stanford-Binet was determined by correlating IQ's on Forms L and M administered to the standardization group within an interval of one week or less. Such reliability coefficients are thus measures of both short-term temporal stability and equivalence of content across the two item samples. An exceptionally thorough analysis of the reliability of this test was carried out with reference to age and IQ level of subjects (McNemar, 1942, Ch. 6). In general, the Stanford-Binet tends to be more reliable for the older than for the younger ages, and for the lower than for the higher IQ's. Thus, at ages 2½ to 5½, the reliability coefficients range from .83 (for IQ 140–149) to .91 (for IQ 60–69); for ages 6 to 13, they range from .91 to .97, respectively, for the same IQ levels; and for ages 14 to 18, the corresponding range of reliability coefficients extends from .95 to .98.

The increasing reliability of scores with increasing age is characteristic of tests in general. It results in part from the better control of conditions that is possible with older subjects (especially in comparison with the preschool ages). Another factor is the slowing down of developmental rate with age. When reliability is measured by retesting, individuals who are undergoing less change are also likely to exhibit less random fluctuation over short periods of time (Pinneau, 1961, Ch. 5).

The higher reliability obtained with lower IQ levels at any given CA, on the other hand, appears to be associated with the specific structural characteristics of the Stanford-Binet. It will be recalled that because of the difference in number of items available at different age levels, each item receives a weight of one month at the lowest levels, a weight of two months at the intermediate levels, and weights of four, five, or six months at the highest levels. This weighting tends to magnify the error of measurement at the upper levels, because the chance passing or failure of a single item makes a larger difference in total score at these levels than it does at lower levels. Since at any given CA, individuals with higher IQ's are tested with higher age levels on the scale, their IQ's will have a larger error of measurement and lower reliability (Pinneau, 1961, Ch. 5). The relationship between IQ level and reliability of the Stanford-Binet is also illustrated graphically in Figure 30, showing the bivariate distribution of IQ's obtained by 7-year-old children on Forms L and M. It will be observed that the individual entries fall close to the diagonal at lower IQ levels and spread farther apart at the higher levels. This indicates closer agreement between L and M IQ's at lower levels and wider discrepancies between them at upper levels. With such a fan-shaped scatter diagram, a single correlation coefficient is misleading. For this reason, separate reliability coefficients have been reported for different portions of the IQ range.

On the whole, the data indicate that the Stanford-Binet is a highly reliable test, most of the reported reliability coefficients for the various age and IQ levels being over .90. Such high reliability coefficients were obtained despite the fact that they were computed separately within each age group. It will be recalled in this connection that all subjects in the

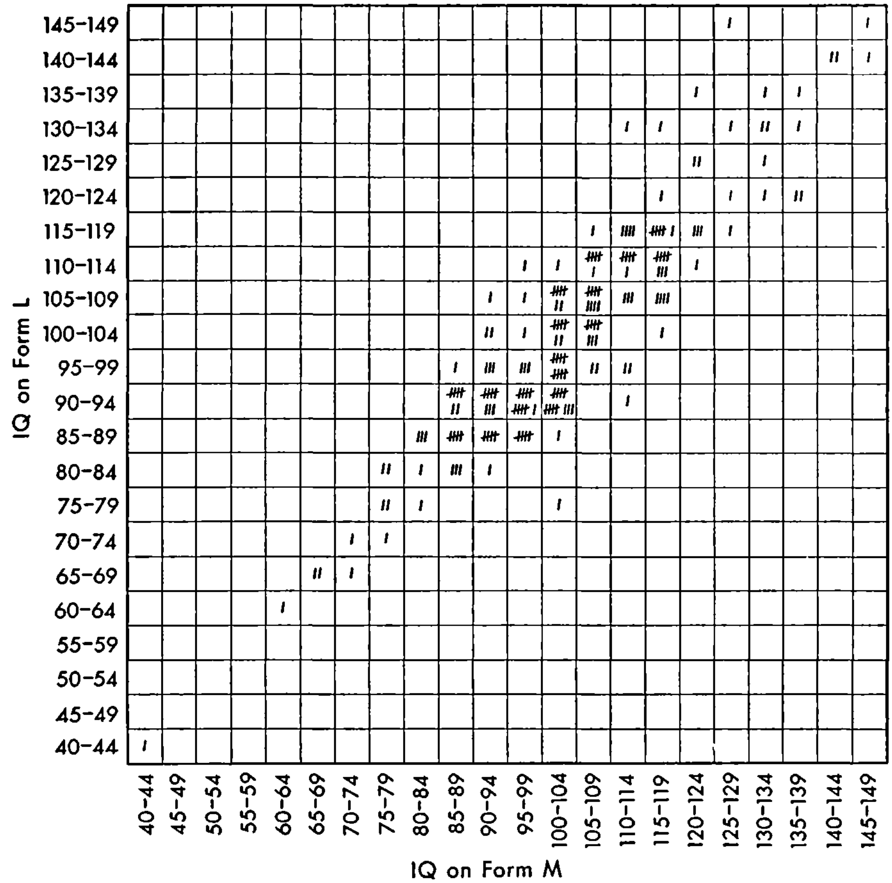


FIG. 30. Parallel-Form Reliability of the Stanford-Binet: Bivariate Distribution of IQ's Obtained by Seven-Year-Old Children on Forms L and M.

(From Terman & Merrill, 1937, p. 45. Reproduced by permission of Houghton Mifflin Company.)

standardization sample were tested within a month of a birthday or half-year birthday. This narrowly restricted age range would tend to produce lower reliability coefficients than found for most tests, which employ more heterogeneous samples. Translated in terms of individual IQ's, a reliability coefficient of .90 and an *SD* of 16 give an error of measurement of approximately 5 IQ points (see Ch. 5). In other words, the chances are about 2:1 that a child's true Stanford-Binet IQ differs by 5 points or less from the IQ obtained in a single testing, and the chances are 95 out of 100 that it varies by no more than 10 points ($5 \times 1.96 = 9.8$). Reflecting the same differences found in the reliability coefficients, these errors of measurement will be somewhat higher for younger than for older children, and somewhat higher for brighter than for duller individuals.

VALIDITY. Some information bearing on the *content validity* of the Stanford-Binet is provided by an examination of the tasks to be performed by the examinee in the various tests. These tasks run the gamut from simple manipulation to abstract reasoning. At the earliest age levels, the tests require chiefly eye-hand coordination, perceptual discrimination, and ability to follow directions, as in block building, stringing beads, comparing lengths, and matching geometric forms. A relatively large number of tests at the lower levels also involve the identification of common objects presented in toy models or in pictures.

Several tests occurring over a wide age range call for practical judgment or common sense. For example, the child is asked, "What should you do if you found on the streets of a city a three-year-old baby that was lost from its parents?" In other tests the examinee is asked to explain why certain practices are commonly followed or certain objects are employed in daily living. A number of tests calling for the interpretation of pictorially or verbally presented situations, or the detection of absurdities in either pictures or brief stories, also seem to fall into this category. Memory tests are found throughout the scale and utilize a wide variety of materials. The individual is required to recall or recognize objects, pictures, geometric designs, bead patterns, digits, sentences, and the content of passages. Several tests of spatial orientation occur at widely scattered levels. These include maze-tracing, paper-folding, paper-cutting, rearrangement of geometric figures, and directional orientation. Skills acquired in school, such as reading and arithmetic, are required for successful performance at the upper year levels.

The most common type of test, especially at the upper age levels, is that employing verbal content. In this category are to be found such well-known tests as vocabulary, analogies, sentence completion, disarranged sentences, defining abstract terms, and interpreting proverbs. Some stress verbal fluency, as in naming unrelated words as rapidly as possible, giving rhymes, or building sentences containing three given words. It should also be noted that many of the tests that are not predominantly verbal in content nevertheless require the understanding of fairly complex verbal instructions. That the scale as a whole is heavily weighted with verbal ability is indicated by the correlations obtained between the 45-word vocabulary test and mental ages on the entire scale. These correlations were found to be .71, .83, .86, and .83 for groups of examinees aged 8, 11, 14, and 18 years, respectively (McNemar, 1942, pp. 139–140;

see also A. J. Edwards, 1963).[5] The correlations are at least as high as those normally found between tests designed to measure the same functions, and they fall within the range of common reliability coefficients.

Insofar as all the functions listed are relevant to what is commonly regarded as "intelligence," the scale may be said to have content validity. The preponderance of verbal content at the upper levels is defended by the test authors on theoretical grounds. Thus, they write:

> At these levels the major intellectual differences between subjects reduce largely to differences in the ability to do conceptual thinking, and facility in dealing with concepts is most readily sampled by the use of verbal tests. Language, essentially, is the shorthand of the higher thought processes, and the level at which this shorthand functions is one of the most important determinants of the level of the processes themselves (Terman & Merrill, 1937, p. 5).

It should be added that clinical psychologists have developed several schemes for classifying Stanford-Binet tests, as aids in the qualitative description of the individual's test performance (see Sattler, 1974, Ch. 10). Pattern analyses of the examinee's successes and failures in different functions may provide helpful clues for further clinical exploration. The results of such analyses, however, should be regarded as tentative and interpreted with caution. Most functions are represented by too few tests to permit reliable measurement; and the coverage of any one function varies widely from one year level to another.

Data on the *criterion-related validity* of the Stanford-Binet, both concurrent and predictive, have been obtained chiefly in terms of academic achievement. Since the publication of the original 1916 Scale, many correlations have been computed between Stanford-Binet IQ and school grades, teachers' ratings, and achievement test scores. Most of these correlations fall between .40 and .75. School progress was likewise found to be related to Stanford-Binet IQ. Children who were accelerated by one or more grades averaged considerably higher in IQ than those at normal age-grade location; and children who were retarded by one or more grades averaged considerably below (McNemar, 1942, Ch. 3).

Like most intelligence tests, the Stanford-Binet correlates highly with performance in nearly all academic courses, but its correlations are highest with the predominantly verbal courses, such as English and history. Correlations with achievement test scores show the same pattern. In a study of high school sophomores, for example, Form L IQ's correlated

---

[5] Since these are part-whole correlations, they are spuriously raised by the inclusion of the vocabulary test in the determination of MA. This effect is slight, however, since the vocabulary test constitutes less than 5 percent of the total number of test items (McNemar, 1942, p. 140).

.73 with Reading Comprehension scores, .54 with Biology scores, and .48 with Geometry scores (Bond, 1940). Correlations in the .50's and .60's have been found with college grades. Among college students, both selective factors and insufficient test ceiling frequently lower the correlations.

There have been relatively few validation studies with the 1960 Form L-M (see Himelstein, 1966). Kennedy, Van de Reit, and White (1963) report a correlation of .69 with total score on the California Achievement Test in a large sample of black elementary school children. Correlations with scores on separate parts of the same battery were: Reading, .68; Arithmetic, .64; and Language, .70.

In interpreting the IQ, it should be borne in mind that the Stanford-Binet—like most so-called intelligence tests—is largely a measure of scholastic aptitude and that it is heavily loaded with verbal functions, especially at the upper levels. Individuals with a language handicap, as well as those whose strongest abilities lie along nonverbal lines, will thus score relatively low on such a test. Similarly, there are undoubtedly a number of fields in which scholastic aptitude and verbal comprehension are not of primary importance. Obviously, to apply any test to situations for which it is inappropriate will only reduce its effectiveness. Because of the common identification of Stanford-Binet IQ with the very concept of intelligence, there has been a tendency to expect too much from this one test.

Data on the *construct validity* of the Stanford-Binet come from many sources. Continuity in the functions measured in the 1916, 1937, and 1960 scales was ensured by retaining in each version only those items that correlated satisfactorily with mental age on the preceding form. Hence, the information that clinicians have accumulated over the years regarding typical behavior of individuals at different MA and IQ levels can be utilized in their interpretation of scores on this scale.

Age differentiation represents the major criterion in the selection of Stanford-Binet items. Thus, there is assurance that the Stanford-Binet measures abilities that increase with age during childhood and adolescence in our culture. In each form, internal consistency was a further criterion for item selection. That there is a good deal of functional homogeneity in the Stanford-Binet, despite the apparent variety of content, is indicated by a mean item-scale correlation of .66 for the 1960 revision. The predominance of verbal functions in the scale is shown by the higher correlation of verbal than nonverbal items with performance on the total scale (Terman & Merrill, 1973, pp. 33–34).

Further data pertaining to construct validity are provided by several independent factor analyses of Stanford-Binet items. If IQ's are to be comparable at different ages, the scale should have approximately the same factorial composition at all age levels. For an unambiguous interpretation of IQ's, moreover, the scale should be highly saturated with a single common factor. The latter point has already been discussed in connection with homogeneity in Chapter 5. If the scores were heavily weighted with two group factors, such as verbal and numerical aptitudes, an IQ of, let us say, 115 obtained by different persons might indicate high verbal ability in one case and high numerical ability in the other.

McNemar (1942, Ch. 9) conducted separate factorial analyses of Stanford-Binet items at 14 age levels, including half-year groups from 2 to 5 and year groups at ages 6, 7, 9, 11, 13, 15, and 18. The number of subjects employed in each analysis varied from 99 to 200, and the number of items ranged from 19 to 35. In each of these analyses, tetrachoric correlations were computed between the items, and the resulting correlations were factor analyzed. By including items from adjacent year levels in more than one analysis, some evidence was obtained regarding the identity of the common factor at different ages. The factor loadings of tests that recur at several age levels provided further data on this point. In general, the results of these analyses indicated that performance on Stanford-Binet items is largely explicable in terms of a single common factor. Evidence of additional group factors was found at a few age levels, but the contribution of these factors was small. It was likewise demonstrated that the common factor found at adjacent age levels was essentially the same, although this conclusion may not apply to more widely separated age levels. In fact, there was some evidence to suggest that the common factor becomes increasingly verbal as the higher ages are approached. The common factor loading of the vocabulary test, for example, rose from .59 at age 6 to .91 at age 18.

Other factor-analytic studies of both the 1937 and the 1960 forms have used statistical techniques designed to bring out more fully the operation of group factors (L. V. Jones, 1949, 1954; Ramsey & Vane, 1970; Sattler, 1974, Ch. 10; Stott & Ball, 1965). Among the factors thus identified were several verbal, memory, reasoning, spatial visualization, and perceptual abilities. In general, the results suggest that there is much in common in the scale as a whole—a characteristic that is largely built into the Stanford-Binet by selecting items that have high correlations with total scores. At the same time, performance is also influenced by a number of special abilities whose composition varies with the age level tested.

WECHSLER ADULT INTELLIGENCE SCALE

The rest of this chapter is concerned with the intelligence scales prepared by David Wechsler. Although administered as individual tests and designed for many of the same uses as the Stanford-Binet, these scales differ in several important ways from the earlier test. Rather than being organized into age levels, all items of a given type are grouped into subtests and arranged in increasing order of difficulty within each subtest. In this respect the Wechsler scales follow the pattern established for group tests, rather than that of the Stanford-Binet. Another characteristic feature of these scales is the inclusion of verbal and performance subtests, from which separate verbal and performance IQ's are computed.

Besides their use as measures of general intelligence, the Wechsler scales have been investigated as a possible aid in psychiatric diagnosis. Beginning with the observation that brain damage, psychotic deterioration, and emotional difficulties may affect some intellectual functions more than others, Wechsler and other clinical psychologists argued that an analysis of the individual's relative performance on different subtests should reveal specific psychiatric disorders. The problems and results pertaining to such a profile analysis of the Wechsler scales will be analyzed in Chapter 16, as an example of the clinical use of tests.

The interest aroused by the Wechsler scales and the extent of their use is attested by some 2,000 publications appearing to date about these scales. In addition to the usual test reviews in the *Mental Measurements Yearbooks,* research pertaining to the Wechsler scales has been surveyed periodically in journals (Guertin et al., 1956, 1962, 1966, 1971; Littell, 1960; Rabin & Guertin, 1951; Zimmerman & Woo-Sam, 1972) and has been summarized in several books (Glasser & Zimmerman, 1967; Matarazzo, 1972; Wechsler, 1958; Zimmerman, Woo-Sam, & Glasser, 1973).

ANTECEDENTS OF THE WAIS. The first form of the Wechsler scales, known as the Wechsler-Bellevue Intelligence Scale, was published in 1939. One of the primary objectives in its preparation was to provide an intelligence test suitable for adults. In first presenting this scale, Wechsler pointed out that previously available intelligence tests had been designed primarily for schoolchildren and had been adapted for adult use by adding more difficult items of the same kinds. The content of such tests was often of little interest to adults. Unless the test items have a certain minimum of face validity, rapport cannot be properly established with adult examinees. Many intelligence test items, written with special reference to the daily activities of the schoolchild, clearly lack face validity for most adults. As Wechsler (1939, p. 17) expressed it, "Asking the ordinary housewife to furnish you with a rhyme to the words, 'day,' 'cat,' and 'mill,' or an ex-army sergeant to give you a sentence with the words, 'boy,' 'river,' 'ball," is not particularly apt to evoke either interest or respect."

The overemphasis on speed in most tests also tends to handicap the older person. Similarly, Wechsler believed that relatively routine manipulation of words received undue weight in the traditional intelligence test. He likewise called attention to the inapplicability of mental age norms to adults, and pointed out that few adults had previously been included in the standardization samples for individual intelligence tests.

It was to meet these various objections that the original Wechsler-Bellevue was developed. In form and content, this scale was closely similar to the more recent Wechsler Adult Intelligence Scale (WAIS) which has now supplanted it. The earlier scale had a number of technical deficiencies, particularly with regard to size and representativeness of normative sample and reliability of subtests, which were largely corrected in the later revision. A further revision is scheduled for 1978.

DESCRIPTION. Published in 1955, the WAIS comprises eleven subtests. Six subtests are grouped into a Verbal Scale and five into a Performance Scale. These subtests are listed and briefly described below, in the order of their administration.

VERBAL SCALE

1. *Information:* 29 questions covering a wide variety of information that adults have presumably had an opportunity to acquire in our culture. An effort was made to avoid specialized or academic knowledge. It might be added that questions of general information have been used for a long time in informal psychiatric examinations to establish the individual's intellectual level and his practical orientation.
2. *Comprehension:* 14 items, in each of which the examinee explains what should be done under certain circumstances, why certain practices are followed, the meaning of proverbs, etc. Designed to measure practical judgment and common sense, this test is similar to the Stanford-Binet Comprehension items; but its specific content was chosen so as to be more consonant with the interests and activities of adults.
3. *Arithmetic:* 14 problems similar to those encountered in elementary school arithmetic. Each problem is orally presented and is to be solved without the use of paper and pencil.
4. *Similarities:* 13 items requiring the subject to say in what way two things are alike.
5. *Digit Span:* Orally presented lists of three to nine digits are to be orally reproduced. In the second part, the examinee must reproduce lists of two to eight digits backwards.
6. *Vocabulary:* 40 words of increasing difficulty are presented both orally and visually. The examinee is asked what each word means.

PERFORMANCE SCALE

7. *Digit Symbol:* This is a version of the familiar code-substitution test which has often been included in nonlanguage intelligence scales. The

key contains 9 symbols paired with the 9 digits. With this key before him, the examinee has 1½ minutes to fill in as many symbols as he can under the numbers on the answer sheet.

8. *Picture Completion:* 21 cards, each containing a picture from which some part is missing. Examinee must tell what is missing from each picture.
9. *Block Design:* This subtest uses a set of cards containing designs in red and white and a set of identical one-inch blocks whose sides are painted red, white, and red-and-white. The examinee is shown one design at a time, which he must reproduce by choosing and assembling the proper blocks (see Fig. 31).

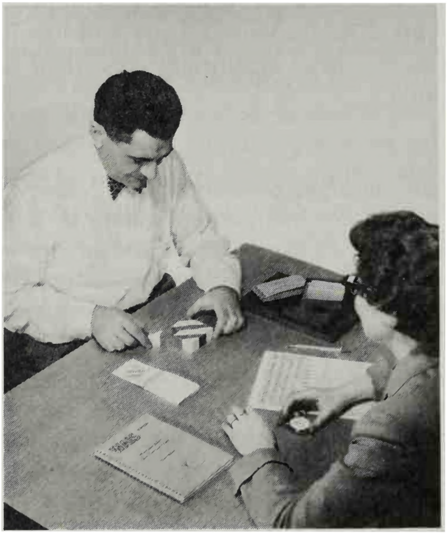


Fig. 31. The Block Design Test of the Wechsler Adult Intelligence Scale.
(Courtesy the Psychological Corporation.)

10. *Picture Arrangement:* Each item consists of a set of cards containing pictures to be rearranged in the proper sequence so as to tell a story. Figure 32 shows one set of cards in the order in which they are presented to the examinee. This set shows the easiest of 8 items making up the test.
11. *Object Assembly:* In each of the four parts of this subtest, cutouts are to be assembled to make a flat picture of a familiar object.

Both speed and correctness of performance are taken into account in scoring Arithmetic, Digit Symbol, Block Design, Picture Arrangement, and Object Assembly.

Since the publication of the original Wechsler-Bellevue scale, a large number of *abbreviated scales* have been proposed. These scales are formed simply by omitting some of the subtests and prorating scores to obtain a Full Scale IQ comparable to the published norms. The fact that several subtest combinations, while effecting considerable saving in time, correlate over .90 with Full Scale IQ's has encouraged the development and use of abbreviated scales for rapid screening purposes. Extensive research has been conducted to identify the most effective combinations of two, three, four, and five subtests in predicting Verbal, Performance, and Full Scale IQ's (Doppelt, 1956; Levy, 1968; Maxwell, 1957; McNemar, 1950; Silverstein, 1970, 1971; Tellegen & Briggs, 1967). A comparative analysis of a single four-subtest combination at different age levels from 18–19 to 75 and over yielded correlations of .95 to .97 with Full Scale IQ's (Doppelt, 1956). Equally close correspondences have been found in several studies of abbreviated scales formed by reducing the number of items within subtests (see Guertin et al., 1966, pp. 388–389; Matarazzo, 1972, pp. 252–255). Much of this research has utilized the WAIS standardization data; but similar studies have been conducted on mental retardates and psychiatric patients (see Matarazzo, 1972, p. 252).

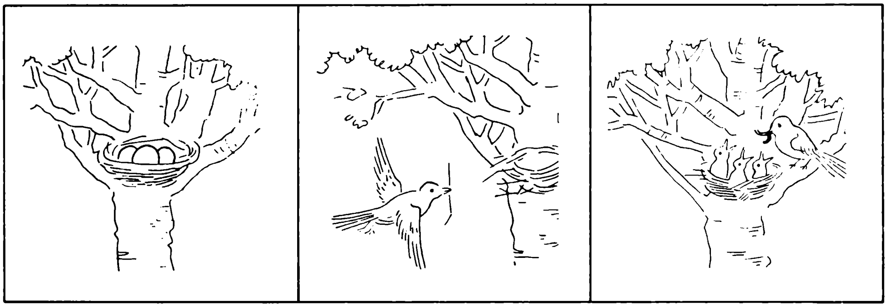


FIG. 32. Easy Item from the WAIS Picture Arrangement Test.

(Reproduced by permission. Copyright © 1955, The Psychological Corporation, New York, N.Y. All rights reserved.)

Although an excessive amount of energy seems to have been expended in assembling and checking short forms of the Wechsler scales, it is probably inadvisable to use such abbreviated versions except as rough screening devices. Many of the qualitative observations made possible by the administration of an individual scale are lost when abbreviated scales are used. Moreover, the assumption that the original Full Scale norms are applicable to prorated total scores on short scales may not always be justified.

NORMS. The WAIS standardization sample was chosen with exceptional care to insure its representativeness. The principal normative sample consisted of 1,700 cases, including an equal number of men and women distributed over seven age levels between 16 and 64 years. Subjects were selected so as to match as closely as possible the proportions in the 1950 United States census with regard to geographical region, urban-rural residence, race (white versus nonwhite), occupational level, and education. At each age level, one man and one woman from an institution for mental retardates were also included. Supplementary norms for older persons were established by testing an "old-age sample" of 475 persons, aged 60 years and over, in a typical midwestern city (Doppelt & Wallace, 1955).

It is admittedly difficult to obtain a representative sample of the population over 60. Although the WAIS sample is probably more nearly representative than any other elderly sample tested prior to that time, there is evidence to suggest that significant regional differences occur in the relative magnitude of Verbal and Performance scores at these age levels (Eisdorfer & Cohen, 1961). Furthermore, the current applicability of norms gathered prior to 1955 is questionable, in view of the rapidly rising educational and cultural level of the population. Especially important in this connection is the rechecking of the age decrement among older persons.

Raw scores on each WAIS subtest are transmuted into standard scores with a mean of 10 and an *SD* of 3. These scaled scores were derived from a reference group of 500 cases which included all persons between the ages of 20 and 34 in the standardization sample. All subtest scores are thus expressed in comparable units and in terms of a fixed reference group. Verbal, Performance, and Full Scale scores are found by adding the scaled scores on the six Verbal subtests, the five Performance subtests, and all eleven subtests, respectively. By reference to appropriate tables provided in the manual, these three scores can be expressed as deviation IQ's with a mean of 100 and an *SD* of 15. Such IQ's, however, are found with reference to the individual's own age group. They therefore show the individual's standing in comparison with persons of his own age level.

In the interpretation of WAIS IQ's, the relative magnitude of IQ's obtained on the Wechsler scales and on other intelligence tests should also be taken into account. It has been repeatedly found that brighter subjects tend to score higher on the Stanford-Binet than on the Wechsler scales, while duller subjects score higher on the Wechsler than on the Stanford-Binet. For example, studies of college freshmen show significantly higher mean IQ's on the Stanford-Binet than on the Wechsler, while the reverse is generally found among the mentally retarded. To some extent, the difference in standard deviation of Wechsler and Stanford-Binet IQ's may account for the differences between the IQ's obtained with the two scales. It will be recalled that the *SD* of the Stanford-Binet IQ is 16, while that of the Wechsler IQ is 15. The discrepancies in individual IQ's, however, are larger than would be expected on the basis of such a difference. Another difference between the two scales is that the Wechsler has less floor and ceiling than the Stanford-Binet and hence does not discriminate as well at the extremes of the IQ range.

The relationship between Stanford-Binet and Wechsler IQ's depends not only on IQ level, but also on age. Other things being equal, older subjects tend to obtain higher IQ's on the Wechsler than on the Stanford-Binet, while the reverse is true of younger subjects. One explanation for such a trend is obviously provided by the use of a declining standard in the computation of the Wechsler IQ's of older persons. On the Stanford-Binet, on the other hand, all adults are evaluated in terms of the average peak age on that scale, viz., 18 years. It is also possible that, since the Stanford-Binet was standardized primarily on children and the Wechsler on adults, the content of the former tends to favor children while that of the latter favors older persons.

RELIABILITY. For each of the eleven subtests, as well as for Verbal, Performance, and Full Scale IQ's, reliability coefficients were computed within the 18–19, 24–34, and 45–54 year samples. These three groups were chosen as being representative of the age range covered by the standardization sample. Odd-even reliability coefficients (corrected for full test length by the Spearman-Brown formula) were employed for every subtest except Digit Span and Digit Symbol. The reliability of Digit Span was estimated from the correlation between Digits Forward and Digits Backward scores. No split-half technique could be utilized with Digit Symbol, which is a highly speeded test. The reliability of this test was therefore determined by parallel-form procedures in a group specially tested with WAIS and Wechsler-Bellevue Digit Symbol subtests.

Full Scale IQ's yielded reliability coefficients of .97 in all three age samples. Verbal IQ's had identical reliabilities of .96 in the three groups, and Performance IQ's had reliabilities of .93 and .94. All three IQ's are thus highly reliable in terms of internal consistency. As might be expected, the individual subtests yield lower reliabilities, ranging from a few coefficients in the .60's found with Digit Span, Picture Arrangement, and Object Assembly, to coefficients as high as .96 for Vocabulary. It is particularly important to consider these subtest reliabilities when evaluating the significance of differences between subtest scores obtained by the same individual, as in profile analysis.

The WAIS manual also reports standard errors of measurement for the three IQ's and for subtest scores. For Verbal IQ, such errors were 3 points in each group, for Performance IQ, just under 4 points, and for Full Scale IQ, 2.60. We could thus conclude, for example, that the chances are roughly 2:1 that an individual's true Verbal IQ falls within 3 points of his obtained Verbal IQ. The above values compare favorably with the 5-point error of measurement found for the Stanford-Binet. It should be remembered, however, that the Stanford-Binet reliabilities were based on parallel forms administered over intervals of one week or less; under such conditions we would anticipate somewhat lower reliability coefficients and greater fluctuation of scores.

VALIDITY. Any discussion of validity of the WAIS must draw on research done with the earlier Wechsler-Bellevue as well. Since all changes introduced in the WAIS represent improvements over the Wechsler-Bellevue (in reliability, ceiling, normative sample, etc.) and since the nature of the test has remained substantially the same, it is reasonable to suppose that validity data obtained on the Wechsler-Bellevue will underestimate rather than overestimate the validity of the WAIS.

The WAIS manual itself contains no validity data, but several aspects of validity are covered in the subsequent books by Wechsler (1958) and by Matarazzo (1972). Wechsler (1958, Ch. 5) argues that the psychological functions tapped by each of the 11 chosen subtests fit the definition of intelligence, that similar tests have been successfully employed in previously developed intelligence scales, and that such tests have proved their worth in clinical experience. The test author himself places the major emphasis on this approach to validity. The treatment is essentially in terms of *content validity*, although it has overtones of construct validity without supporting data. Much of the discussion in Matarazzo's book is of the same nature, dealing with the construct of global intelligence, but having only a tenuous relation to the evaluation of the WAIS as a measuring instrument.

Some empirical data on *concurrent criterion-related validity* are summarized in the two books (Matarazzo, 1972, p. 284; Wechsler, 1958, Ch. 14). These data include mean IQ differences among various educational and occupational groups, as well as a few correlations with job-performance ratings and academic grades. Most group differences, though small, are in the expected directions. Persons in white-collar jobs of different kinds and levels averaged higher in Verbal than in Performance IQ, but skilled workers averaged higher in Performance than in Verbal. In studies of industrial executives and psychiatric residents, Verbal IQ correlated in the .30's with overall performance ratings. Both groups, of course, were already selected in terms of the abilities measured by these tests. Correlations in the .40's and .50's have been found between Verbal IQ and college or engineering school grades. In all these groups, the Verbal Scale yielded somewhat higher correlations than the Full Scale; correlations with the Performance Scale were much lower. Even the correlations with the Verbal Scale, however, were not appreciably higher than those obtained with the Stanford-Binet and with well-known group tests. In studies of mental retardates, WAIS IQ's have proved to be satisfactory predictors of institutional release rate and subsequent work adjustment (see Guertin et al., 1966).

The Wechsler scales have been repeatedly correlated with the Stanford-Binet as well as with other well-known tests of intelligence (Guertin et al., 1971; Matarazzo, 1972; Wechsler, 1958). Correlations with the Stanford-Binet in unselected adolescent or adult groups and among mental retardates cluster around .80. Within more homogeneous samples, such as college students, the correlations tend to be considerably lower. Group tests yield somewhat lower correlations with the Wechsler scales, although such correlations vary widely as a function of the particular test and the nature and heterogeneity of the sample. For both Stanford-Binet and group scales, correlations are nearly always higher with the Wechsler Verbal Scale than with the Full Scale, while correlations with the Performance Scale are much lower than either. On the other hand, Performance IQ's correlate more highly than Verbal IQ's with tests of spatial abilities. For example, a correlation of .72 was found between Performance IQ and the Minnesota Paper Form Board Test in a group of 16-year-old boys and girls (Janke & Havighurst, 1945). In other studies, Performance IQ's correlated .70 with Raven's Progressive Matrices (Hall, 1957) and .35 with the Bennett Mechanical Comprehension Test (Wechsler, 1958, p. 228).

Of some relevance to the *construct validity* of the Wechsler scales are the intercorrelations of subtests and of Verbal and Performance IQ's, as well as factorial analyses of the scales. In the process of standardizing the WAIS, intercorrelations of Verbal and Performance Scales and of the 11 subtests were computed on the same three age groups on which reliability coefficients had been found, namely, 18–19, 25–34, and 45–54. Verbal and Performance Scale scores correlated .77, .77, and .81, respectively, in these three groups. Intercorrelations of separate subtests were also similar in the three age groups, running higher among Verbal than among Performance subtests. Correlations between Verbal and Performance subtests, although still lower on the whole, were substantial. For example, in the 25–34 year group, correlations among Verbal subtests ranged from .40 to .81, among Performance subtests from .44 to .62, and between Performance and Verbal subtests from .30 to .67. Both individual subtest correlations and correlations between total Verbal and Performance Scale scores suggest that the two scales have much in common and that the allocation of tests to one or the other scale may be somewhat arbitrary.

Factorial analyses of the Wechsler scales have been conducted with a variety of subjects ranging from eighth-grade pupils to the old-age standardization sample (aged 60–75+) and including both normal and abnormal groups. They have also employed different statistical procedures and have approached the analysis from different points of view. Some have been directly concerned with age changes in the factorial organization of the Wechsler subtests, but the findings of different investigators are inconsistent in this regard. As an example, we may examine the factorial analyses of the WAIS conducted by J. Cohen (1957a, 1957b) with the intercorrelations of subtests obtained on four age groups in the standardization sample (18–19, 25–34, 45–54, and 60–75+). The major results of this study are in line with those of other investigations using comparable procedures, as well as with the findings of later studies by Cohen and his associates on different populations (see Guertin et al., 1962, 1966).

That all 11 subtests have much in common was demonstrated in Cohen's study by the presence of a single general factor that accounted for about 50 percent of the total variance of the battery. In addition, three major group factors were identified. One was a *verbal comprehension* factor, with large weights in the Vocabulary, Information, Comprehension, and Similarities subtests. A *perceptual organization* factor was found chiefly in Block Design and Object Assembly. This factor may actually represent a combination of the perceptual speed and spatial visualization factors repeatedly found in factorial analyses of aptitude tests. The results of an earlier investigation by P. C. Davis (1956), in which "reference tests" measuring various factors were included with the Wechsler subtests, support this composite interpretation of the perceptual organization factor.

The third major group factor identified by Cohen was described as a *memory* factor. Found principally in Arithmetic and Digit Span, it apparently includes both immediate rote memory for new material and recall of previously learned material. Ability to concentrate and to resist distraction may be involved in this factor. Of special interest is the finding that the memory factor increased sharply in prominence in the old-age sample. At that age level it had significant loadings, not only in Arithmetic and Digit Span, but also in Vocabulary, Information, Comprehension, and Digit Symbol. Cohen points out that during senescence memory begins to deteriorate at different ages and rates in different persons. Individual differences in memory thus come to play a more prominent part in intellectual functioning than had been true at earlier ages. Many of the

WAIS subtests require memory at all ages. Until differential deterioration sets in, however, individual differences in the retentive ability required in most of the subtests are insignificant.

WECHSLER INTELLIGENCE SCALE FOR CHILDREN

DESCRIPTION. The Wechsler Intelligence Scale for Children (WISC) was first prepared as a downward extension of the original Wechsler-Bellevue (Seashore, Wesman, & Doppelt, 1950). Many items were taken directly from the adult test; and easier items of the same types were added to each test. A revised edition, WISC-R, was published in 1974. The WISC-R consists of twelve subtests, of which two are used only as alternates or as supplementary tests if time permits. The materials used in administering the WISC-R are pictured in Figure 33; and Figure 34 shows a child working on one of the easier items of the Object Assembly

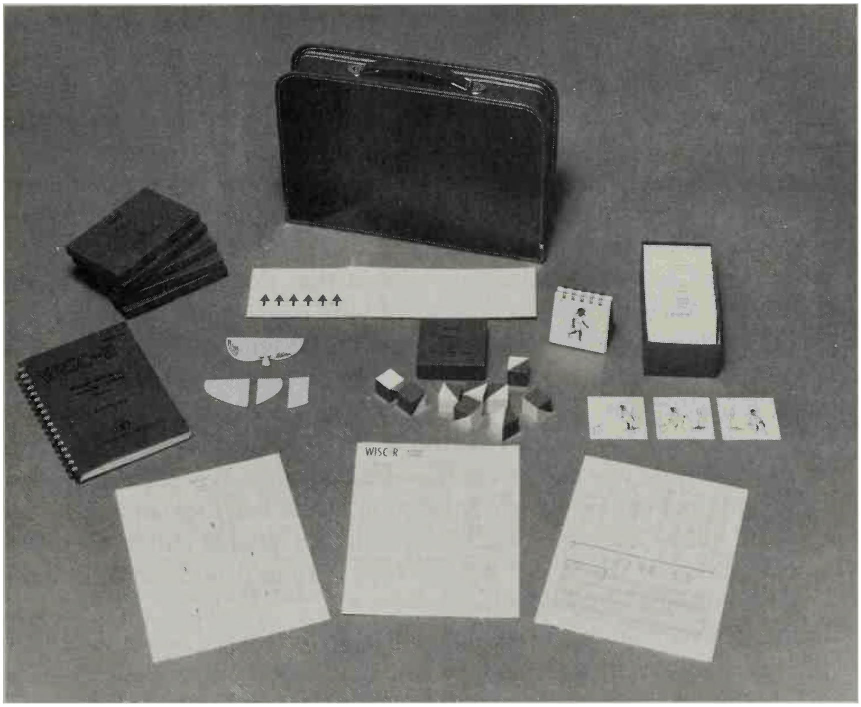


FIG. 33. Materials Used with the Wechsler Intelligence Scale for Children—Revised.

(Courtesy The Psychological Corporation.)

subtest. As in the other Wechsler scales, the subtests are grouped into a Verbal and a Performance scale as follows:

VERBAL SCALE	PERFORMANCE SCALE
1. Information	2. Picture Completion
3. Similarities	4. Picture Arrangement
5. Arithmetic	6. Block Design
7. Vocabulary	8. Object Assembly
9. Comprehension	10. Coding (or Mazes)
(Digit Span)	

The numbers correspond to the order in which the subtests are administered. Unlike the procedure followed in the WAIS and the earlier WISC, the Verbal and Performance subtests in the WISC-R are administered in alternating order. The Mazes subtest, which requires more time, may be substituted for Coding if the examiner so prefers. Any other substitution, including the substitution of Mazes for any other subtest and the substitution of Digit Span for any of the Verbal subtests, should be made only if one of the regular subtests must be omitted because of special handicaps or accidental disruption of testing procedure. The supplementary tests may always be administered in addition to the regular battery and their inclusion is advised because of the qualitative and diagnostic information provided. In such cases, however, their scores are not used in finding IQ's.

With regard to content, the only subtest that does not appear in the adult scale is Mazes. This test consists of nine paper-and-pencil mazes of increasing difficulty, to be completed within designated time limits and scored in terms of errors. The Coding subtest corresponds to the Digit Symbol subtest of the WAIS, with an easier part added. The remaining subtests represent downward extensions of the adult tests. The development of the WISC was somewhat paradoxical, since Wechsler embarked upon his original enterprise partly because of the need for an adult scale that would *not* be a mere upward extension of available children's scales. The first edition of the WISC was, in fact, criticized because its content was not sufficiently child-oriented.

In the revised edition (WISC-R), special efforts were made to replace or modify adult-oriented items so as to bring their content closer to common childhood experiences. In the Arithmetic subtest, for instance, "cigars" was changed to "candy bars" and items about a taxi and a card game were replaced. Other changes included the elimination of items that might be differentially familiar to particular groups of children and the inclusion of more female and black subjects in the pictorial content

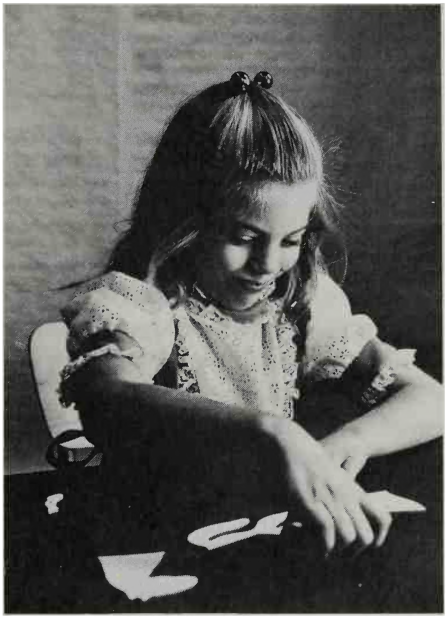


Fig. 34. The Object Assembly Test of the Wechsler Intelligence Scale for Children—Revised.

(Courtesy The Psychological Corporation.)

of the subtests. Several of the subtests were lengthened in order to increase reliability; and improvements were introduced in administration and scoring procedures.

As in the case of the WAIS, there has been considerable experimentation with abbreviated scales of the WISC. The correlations of these short forms with Full Scale IQ's run lower than in the WAIS. With scales consisting of five or six subtests from both Verbal and Performance sets, correlations in the .80's have been found with Full Scale IQ's. It should be noted, however, that these data were obtained with the earlier form of WISC. With the increased length and improved reliability of the WISC-R subtests, the correlations of abbreviated scales with Full Scale IQ's will undoubtedly be higher. Using the WISC standardization data and a procedure that takes subtest reliability into account, Silverstein (1970) identified the ten best combinations of two, three, four, and five WISC subtests. A widely used two-test combination consists of Vocabulary and Block Design. The same cautions mentioned regarding the use of WAIS short forms can be repeated here.

NORMS. The treatment of scores on the WISC-R follows the same procedures used in the adult scale, with minor differences. Raw scores on each subtest are first transmuted into normalized standard scores *within the child's own age group.* Tables of such scaled scores are provided for every four-month interval between the ages of 6-0 and 16-11 years. As in the adult scales, the subtest scaled scores are expressed in terms of a distribution with a mean of 10 and an *SD* of 3 points. The scaled subtest scores are added and converted into a deviation IQ with a mean of 100 and an *SD* of 15. Verbal, Performance, and Full Scale IQ's can be found by the same method.

Although a mental age is not needed to compute a deviation IQ, the WISC-R provides data for interpreting performance on individual subtests in terms of age norms. For each subtest, the manual gives the mean raw score found in the standardization sample for each age from 6-2 to 16-10 at intervals of four months. A child's *test age* can be found by looking up the age corresponding to his raw score. A mean or median test age on the entire scale can be computed if desired.

The standardization sample for the WISC-R included 100 boys and 100 girls at each year of age from 6½ through 16½, giving a total of 2200 cases. Each child was tested within six weeks of his midyear. For example, the 8-year-olds ranged from 8-years 4-months 15-days to 8-years 7-months 15-days. The sample was stratified on the basis of the 1970 U.S. Census with respect to geographic region, urban-rural residence, occupation of head of household, and race (white-nonwhite). Bilinguals were included if they could speak and understand English. Institutionalized mental retardates and children with severe emotional disorders were excluded. Testing was conducted in 32 states (including Hawaii) and Washington, D.C. In many respects, the WISC-R standardization sample is more nearly representative of the U.S. population within the designated age limits than is any other sample employed in standardizing individual tests.

Unlike the earlier WISC and the WAIS, the WISC-R yields mean IQ's that are very close to Stanford-Binet IQ's (1972 norms). The increased similarity in IQ may be attributed in part to the closeness in time when the norms were gathered for the two scales, and in part to the previously mentioned improvements in content, administration, and scoring procedures in the WISC-R.

RELIABILITY. Both split-half and retest coefficients were computed for WISC-R subtests,[6] as well as for Verbal, Performance, and Full Scale IQ's. Odd-even reliabilities were computed separately within each of the 11 age groups; retest coefficients were found within three age groups (6½–

---

[6] Only retest coefficients are reported for Coding and Digit Span, for which split-half coefficients would be inappropriate.

7½, 10½–11½, 14½–15½), the tests being readministered after approximately a one-month interval. Average split-half reliabilities for Verbal, Performance, and Full Scale IQ's were .94, .90, and .96. The corresponding retest coefficients were .93, .90, and .95. Some practice effect was observed on the retest, with mean gains of 3½ IQ points on the Verbal Scale, 9½ on the Performance Scale, and 7 on the Full Scale. Such a practice effect should be taken into account when retesting children after short intervals of time.

Subtest reliabilities are generally satisfactory and higher than in the earlier form. Split-half reliabilities averaged across age groups range from .70 to .86; and mean retest coefficients range from .65 to .88. A noteworthy feature of the manual is the inclusion of tables giving the standard error of measurement for subtests and for Verbal, Performance, and Full Scale IQ's within each age group, as well as the minimum difference between scores required for significance at specified levels. In comparisons of Verbal and Performance IQ's, a difference of 11 or 12 points is significant at the .05 level. The standard error of the Full Scale IQ is approximately 3 points. Thus the chances are 95 out of 100 that a child's true WISC-R IQ differs from his obtained IQ by no more than ±6 points ($3 \times 1.96 = 5.88$).

VALIDITY. No discussion of validity is included in the WISC-R manual. To be sure, normative tables of standard score equivalents for each subtest provide evidence of age differentiation, but no evaluation of the data in terms of this criterion is given. A number of independent investigators have found concurrent validity coefficients between the earlier WISC and achievement tests or other academic criteria of intelligence clustering between .50 and .60 (Littell, 1960; Zimmerman & Woo-Sam, 1972). As would be expected, the Verbal Scale tended to correlate higher than the Performance Scale with such criteria. When children in the WISC standardization sample were classified according to father's occupational level, the usual hierarchy of mean IQ's was found (Seashore, Wesman, & Doppelt, 1950). The difference tended to be slightly larger in Verbal than in Performance IQ and to decrease somewhat with age, possibly because of exposure to relatively uniform education (Estes, 1953, 1955).

The WISC-R manual reports correlations with 1972 Stanford-Binet IQ's within homogeneous age groups. The mean correlation with Full Scale IQ is .73. The Verbal Scale again correlates more highly with the Stanford-Binet than does the Performance Scale (.71 versus .60). Among the subtests, Vocabulary yields the highest mean correlation (.69) and Coding the lowest (.26).

Additional information provided in the WISC-R manual includes intercorrelations among the individual subtests, as well as the correlation of each subtest with Verbal, Performance, and Full Scale scores, and of these three composite scores with each other. All correlations are given separately for the 200 cases in each of the 11 age groups in the standardization sample. The correlations between total Verbal and Performance scores ranged from .60 to .73 within age groups, averaging .67. Thus the two parts of the scale have much in common, although the correlations between them are low enough to justify the retention of the separate scores.

Factorial analyses of the earlier WISC subtests identified factors quite similar to those found in the adult scales, namely general, verbal comprehension, perceptual-spatial, and memory (or freedom from distractibility) factors (see Littell, 1960; Zimmerman & Woo-Sam, 1972). In a more recent study (Silverstein, 1973), the WISC subtests were factor analyzed separately in groups of 505 English-speaking white, 318 black, and 487 Mexican-American children aged 6 to 11 years. The results revealed a verbal comprehension factor having substantial correlations with the five verbal tests; and a perceptual organization factor having substantial correlations with Block Design and Object Assembly. A major finding of this study was the similarity of factor structure across the three ethnic groups, suggesting that the tests measure the same abilities in these groups. A factor analysis of the WISC-R scores of the standardization sample at 11 age levels between 6½ and 16½ years yielded clear evidence of three major factors at each age level (Kaufman, 1975a). These factors corresponded closely to the previously described factors of verbal comprehension, perceptual organization, and freedom from distractibility.

WECHSLER PRESCHOOL AND PRIMARY SCALE OF INTELLIGENCE

DESCRIPTION. In more than one sense, the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) is the baby of the series. Published in 1967, this scale is designed for ages 4 to 6½ years. The scale includes 11 subtests, only 10 of which are used in finding the IQ. Eight of the subtests are downward extensions and adaptations of WISC subtests; the other three were newly constructed to replace WISC subtests that proved unsuitable for a variety of reasons. As in the WISC and WAIS, the subtests are grouped into a Verbal and a Performance scale, from which Verbal, Performance, and Full Scale IQ's are found. As in WISC-R, the administration of Verbal and Performance subtests is alternated in order to enhance variety and help to maintain the young child's interest and cooperation. Total testing time ranges from 50 to 75 minutes, in one or two testing sessions.

In the following list, the new subtests have been starred:

VERBAL SCALE
Information
Vocabulary
Arithmetic
Similarities
Comprehension
*Sentences (Supplementary Test)

PERFORMANCE SCALE
*Animal House
Picture Completion
Mazes
*Geometric Design
Block Design

"Sentences" is a memory test, substituted for the WISC Digit Span. The child repeats each sentence immediately after oral presentation by the examiner. This test can be used as an alternate for one of the other verbal tests; or it can be administered as an additional test to provide further information about the child, in which case it is not included in the total score in calculating the IQ. "Animal House" is basically similar to the WAIS Digit Symbol and the WISC Coding test. A key at the top of the board has pictures of dog, chicken, fish, and cat, each with a differently colored cylinder (its "house") under it. The child is to insert the correctly colored cylinder in the hole beneath each animal on the board (see Fig.

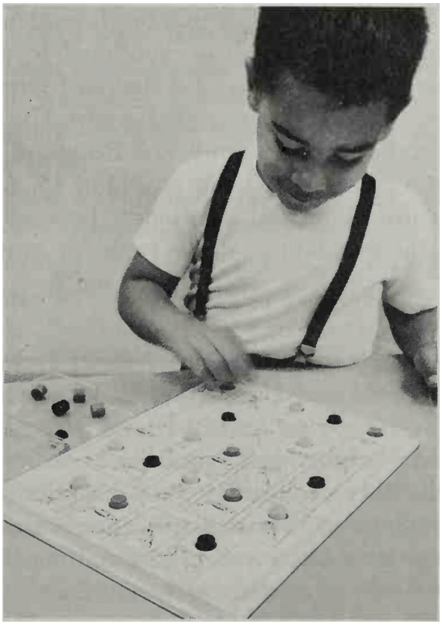


Fig. 35. The Animal House Test of the Wechsler Preschool and Primary Scale of Intelligence.

(Courtesy The Psychological Corporation.)

35). Time, errors, and omissions determine the score. "Geometric Design" requires the copying of 10 simple designs with a colored pencil.

The possibility of using short forms seems to have aroused as much interest for WPPSI as it had for WAIS and WISC. Some of the same investigators have been concerned with the derivation of such abbreviated scales at all three levels, notably Silverstein (1968a, 1968b, 1970, 1971). In a particularly well-designed study, Kaufman (1972) constructed a short form consisting of two Verbal subtests (Arithmetic and Comprehension) and two Performance subtests (Block Design and Picture Completion). Within individual age levels, this battery yielded reliability coefficients ranging from .91 to .94 and correlations with Full Scale IQ ranging from .89 to .92. Half the WPPSI standardization sample of 1,200 cases was used in the selection of the tests; the other half was used to cross validate the resulting battery. Kaufman reiterates the customary caution to use the short form only for screening purposes when time does not permit the administration of the entire scale.

NORMS. The WPPSI was standardized on a national sample of 1,200 children—100 boys and 100 girls in each of six half-year age groups from 4 to 6½. Children were tested within six weeks of the required birthday or midyear date. The sample was stratified against 1960 census data with reference to geographical region, urban-rural residence, proportion of whites and nonwhites, and father's occupational level. Raw scores on each subtest are converted to normalized standard scores with a mean of 10 and an *SD* of 3 within each quarter-year group. The sums of the scaled scores on the Verbal, Performance, and Full Scale are then converted to deviation IQ's with a mean of 100 and an *SD* of 15. Although Wechsler argues against the use of mental age scores because of their possible misinterpretations, the manual provides a table for converting raw scores on each subtest to "test ages" in quarter-year units.

RELIABILITY. For every subtest except Animal House, reliability was found by correlating odd and even scores and applying the Spearman-Brown formula. Since scores on Animal House depend to a considerable extent on speed, its reliability was found by a retest at the end of the testing session. Reliability coefficients were computed separately within each half-year age group in the standardization sample. While varying with subtest and age level, these reliabilities fell mostly in the .80's. Reliability of the Full Scale IQ varied between .92 and .94; for Verbal IQ, it varied between .87 and .90; and for Performance IQ, between .84 and .91. Standard errors of measurement are also provided in the manual, as well as tables for evaluating the significance of the difference between scores. From these data it is suggested that a difference of 15 points or more between Verbal and Performance IQ's is sufficiently important to be investigated.

Stability over time was checked in a group of 50 kindergarten children retested after an average interval of 11 weeks. Under these conditions, reliability of Full Scale IQ was .92; for Verbal IQ it was .86; and for Performance IQ, .89.

VALIDITY. As is true of the other two Wechsler scales, the WPSSI manual contains no section labeled "validity," although it does provide some data of tangential relevance to the validity of the instrument. Intercorrelations of the 11 subtests within each age level in the standardization sample fall largely between .40 and .60. Correlations between Verbal and Performance subtests are nearly as high as those within each scale. The overlap between the two scales is also indicated by an average correlation of .66 between Verbal and Performance IQ's.

The manual reports a correlation of .75 with Stanford-Binet IQ in a group of 98 children aged 5 to 6 years. As in the case of the WISC, the Stanford-Binet correlates higher with the Verbal IQ (.76) than with the Performance IQ (.56). This finding was corroborated in subsequent studies by other investigators working with a variety of groups. In thirteen studies surveyed by Sattler (1974, p. 209), median correlations of WPSSI with Stanford-Binet IQ were .82, .81, and .67 for Full Scale, Verbal, and Performance IQ's, respectively. Correlations have also been found with a number of other general ability tests (for references, see Sattler, 1974, p. 210 and Appendix B-9). Data on predictive validity are meager (Kaufman, 1973a).

A carefully designed reanalysis of the standardization sample of 1,200 cases (Kaufman, 1973b) provided information on the relation of WPSSI scores to socioeconomic status (as indicated by father's occupational level), urban versus rural residence, and geographic region. For each of these three variables, WPSSI Verbal, Performance, and Full Scale IQ's were compared between samples matched on all the other stratification variables (including the other two variables under investigation plus sex, age, and color).

Socioeconomic status yielded significant differences only at the extremes of the distribution. Children with fathers in the professional and technical categories averaged significantly higher than all other groups (Mean IQ = 110); and children whose fathers were in the unskilled category averaged significantly lower than all other groups (Mean IQ = 92.1). Geographic region showed no clear relation to scores. No significant differences was found between matched urban and rural samples, unlike earlier studies with the WISC (Seashore, Wesman, & Doppelt, 1950) and the Stanford-Binet (McNemar, 1942). The investigator attributes the discrepancy principally to the contribution of other variables, which were controlled in this study but not in the earlier studies. Another major difference, however, is in the time when the data were gathered. The intervening 25 or 30 years have witnessed marked changes both in population movements between urban and rural environments and in the educational and cultural facilities available in those environments. It is reasonable to expect that such sociocultural changes could have cancelled out the earlier differences in intelligence test performance characterizing children in these two types of environment.

Since the publication of the WPSSI, several investigators have factor analyzed the subtest scores on samples of diverse populations (see Sattler, 1974, pp. 227–230). A comprehensive study (Hollenbeck & Kaufman, 1973) applied several factor-analytic techniques to three separate age groups in the WPSSI standardization sample. The results provided consistent evidence of a general factor in the battery as a whole, together with two broad group factors: a verbal factor with substantial loadings in the six verbal subtests in each age group; and a performance factor with substantial loadings in the five performance tests in the two older groups and somewhat lower but still appreciable loadings in the youngest group (ages 4 to 4½ years). The separation of the two group factors was generally less clear-cut in the youngest group, a finding that is in line with much of the earlier research on the organization of abilities in young children. When subtest scores were factor analyzed separately for black and white children in the standardization sample, the results in both groups were closely similar to those obtained in the total sample (Kaufman & Hollenbeck, 1974).

* * *

CONCLUDING REMARKS ON THE WECHSLER SCALES. The currently available forms of the three Wechsler scales reflect an increasing level of sophistication and experience in test construction, corresponding to the decade when they were developed: WAIS (1955), WPSSI (1967), WISC-R (1974). In comparison with other individually administered tests, their principal strengths stem from the size and representativeness of the standardization samples, particularly for adult and preschool populations, and the technical quality of their test construction procedures. The treatment of reliability and errors of measurement is especially commendable. The weakest feature of all three scales is the dearth of empirical data on validity. The factor-analytic studies contribute to a clarification of the constructs in terms of which performance on the Wechsler scales may be described; but even these studies would have been more informative if they had included more indices of behavior external to the scales themselves.

CHAPTER 10

Tests for Special Populations

THE TESTS brought together in this chapter include both individual and group scales. They were developed primarily for use with persons who cannot be properly or adequately examined with traditional instruments, such as the individual scales described in the preceding chapter or the typical group tests to be considered in the next chapter. Historically, the kinds of tests surveyed in this chapter were designated as performance, nonlanguage, or nonverbal.

Performance tests, on the whole, involve the manipulation of objects, with a minimal use of paper and pencil. *Nonlanguage tests* require no language on the part of either examiner or examinee. The instructions for these tests can be given by demonstration, gesture, and pantomime, without the use of oral or written language. A prototype of nonlanguage group tests was the Army Examination Beta, developed for testing foreign-speaking and illiterate recruits during World War I (Yerkes, 1921). Revisions of this test were subsequently prepared for civilian use. For most testing purposes, it is not necessary to eliminate all language from test administration, since the examinees usually have some knowledge of a common language. Moreover, short, simple instructions can usually be translated or given successively in two languages without appreciably altering the nature or difficulty of the test. None of these tests, however, requires the examinee himself to use either written or spoken language.

Still another related category is that of *nonverbal tests,* more properly designated as nonreading tests. Most tests for primary school and preschool children fall into this category, as do tests for illiterates and nonreaders at any age level. While requiring no reading or writing, these tests make extensive use of oral instructions and communication on the part of the examiner. Moreover, they frequently measure verbal comprehension, such as vocabulary and the understanding of sentences and short paragraphs, through the use of pictorial material supplemented with oral instructions to accompany each item. Unlike the nonlanguage tests, they would thus be unsuited for foreign-speaking or deaf persons.

Although the traditional categories of performance, nonlanguage, and nonverbal tests contribute to an understanding of the purposes that different tests may serve, the distinctions have become somewhat blurred as more and more test batteries were developed that cut across these categories. The combination of verbal and performance tests in the Wechsler scales is a classical example.

In the present chapter, tests have been classified, not in terms of their content or administrative procedures, but with reference to their principal uses. Three major categories can be recognized from this viewpoint: tests for the infant and preschool level; tests suitable for persons with diverse sensory and motor handicaps; and tests designed for use across cultures or subcultures. Such a classification must remain flexible, however, since several of the tests have proved useful in more than one context. This is especially true of some of the instruments originally designed for cross-cultural testing, which are now more commonly used in clinical testing.

Finally, although some of the tests covered in this chapter were designed as group tests, they are frequently administered individually. A few are widely used in clinical testing to supplement the usual type of intelligence test and thus provide a fuller picture of the individual's intellectual functioning. Several permit the kind of qualitative observations associated with individual testing and may require considerable clinical sophistication for a detailed interpretation of test performance. On the whole, they are closer to the individual tests illustrated in Chapter 9 than to the group tests to be surveyed in Chapter 11.

INFANT AND PRESCHOOL TESTING

All tests designed for infants and preschool children require individual administration. Some kindergarten children can be tested in small groups with the types of tests constructed for the primary grades. In general, however, group tests are not applicable until the child has reached school age. Most tests for children below the age of 6 are either performance or oral tests. A few involve rudimentary manipulation of paper and pencil.

It is customary to subdivide the first five years of life into the infant period and the preschool period. The first extends from birth to the age of approximately 18 months; the second, from 18 to 60 months. From the viewpoint of test administration, it should be noted that the infant must be tested while he is either lying down or supported on a person's lap. Speech is of little use in giving test instructions, although the child's own language development provides relevant data. Many of the tests deal with sensorimotor development, as illustrated by the infant's ability to lift his head, turn over, reach for and grasp objects, and follow a moving object with his eyes. The preschool child, on the other hand, can walk, sit at a table, use his hands in manipulating test objects, and communicate by language. At the preschool level, the child is also much more responsive to the examiner as a person, whereas for the infant the examiner serves primarily as a means of providing stimulus objects. Preschool testing is a more highly interpersonal process—a feature that augments both the opportunities and the difficulties presented by the test situation.

The proper psychological examination of young children requires coverage of a broad spectrum of behavior, including motor and social as well as cognitive traits. This orientation is reflected in some of the pioneer developmental scales to be discussed in the next section; it has also been reaffirmed in recent analyses of the field by specialists in early childhood education (Anderson & Messick, 1974).

DEVELOPMENTAL SCALES. Following a series of longitudinal studies of the normal course of behavior development in the infant and preschool child, Gesell and his associates at Yale prepared the Gesell Developmental Schedules (Gesell & Amatruda, 1947). These schedules cover four major areas of behavior: motor, adaptive, language, and personal-social. They provide a standardized procedure for observing and evaluating the course of behavior development in the child's daily life. Although a few may be properly described as tests, most of the items in these schedules are purely observational. Data are obtained through the direct observation of the child's responses to standard toys and other stimulus objects and are supplemented by information provided by the mother. In evaluating the child's responses, the examiner is aided by very detailed verbal descriptions of the behavior typical of different age levels, together with drawings such as those reproduced in Figure 36. While extending from the age of 4 weeks to 6 years, the Gesell schedules typify the approach followed in infant testing. Items from these schedules have been incorporated in several other developmental scales designed for the infant level.

Although both observational and scoring procedures are less highly standardized in the Gesell schedules than in the usual psychological test, there is evidence that, with adequate training, examiner reliabilities over .95 can be attained (Knobloch & Pasamanick, 1960). In general, these schedules may be regarded as a refinement and elaboration of the qualitative observations routinely made by pediatricians and other specialists concerned with infant development. They appear to be most useful as a supplement to medical examinations for the identification of neurological defect and organically caused behavioral abnormalities in early life (Donofrio, 1965; Knobloch & Pasamanick, 1960).

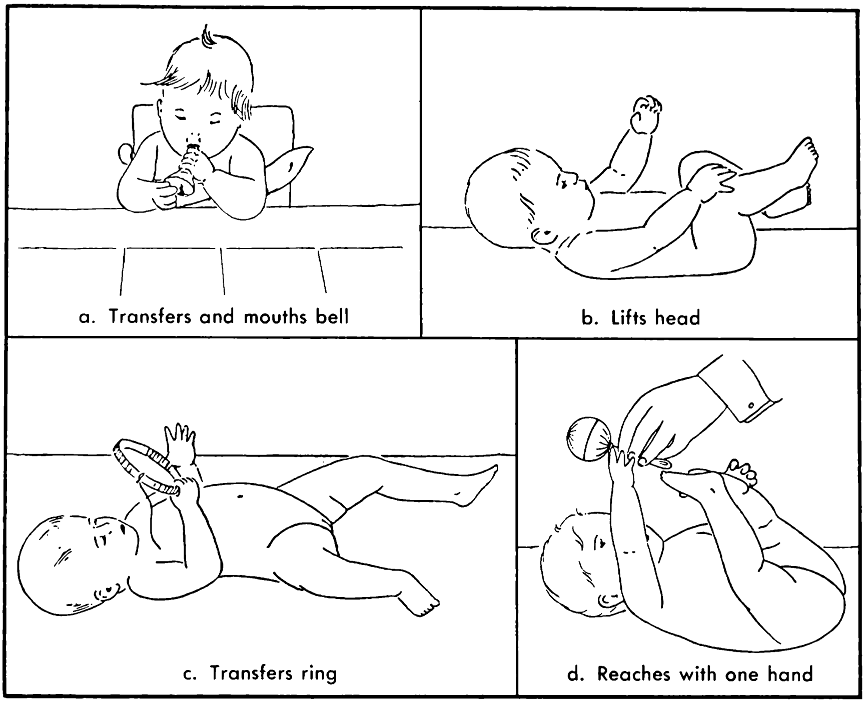

a. Transfers and mouths bell
b. Lifts head
c. Transfers ring
d. Reaches with one hand

FIG. 36. Drawings Employed with the Gesell Developmental Schedules to Illustrate Typical Behavior at 28 Weeks of Age.

(From *Developmental Diagnosis*, by Arnold Gesell and Catherine S. Amatruda. Copyright 1941, 1947, by Arnold Gesell. By permission of Paul B. Hoeber, Inc., publisher.)

Another variety of developmental scale is more restricted in its behavioral coverage but extends over a much wider age range. The prototypes of such scales are the Oseretsky Tests of Motor Proficiency and the Vineland Social Maturity Scale. Although extending well beyond the preschool ages, these scales are relevant to the present discussion because of certain similarities to the Gesell scales in content and in general approach. They are also more suitable for use at the lower age and intellectual levels than at higher levels.

The Oseretsky Tests of Motor Proficiency were originally published in Russia in 1923. They were subsequently translated into several languages and used in a number of European countries. In 1946, Doll (1946), then Director of Research for the Vineland Training School, sponsored and edited an English translation of the Portuguese adaptation of these tests. A scale of motor development is especially useful in testing the mentally retarded, who are also frequently retarded in motor functions. Other applications of the Oseretsky tests are found in the testing of children with motor handicaps, minimal brain dysfunction, or learning disabilities, particularly in connection with the administration of individualized training programs.[1] The original Oseretsky tests extended from 4 to 16 years, the tests being arranged into year levels as in the Stanford-Binet. The Oseretsky scale was designed to cover all major types of motor behavior, from postural reactions and gross bodily movements to finger coordination and control of facial muscles. Administration of these tests requires only simple and easily obtainable materials, such as matchsticks, wooden spools, thread, paper, rope, boxes, and rubber ball. Directions are given orally and by demonstration.

In 1955, the Lincoln-Oseretsky Motor Development Scale (Sloan, 1955) was issued as a revision and restandardization of the Oseretsky tests with simplified instructions and improved scoring procedures. Covering only ages 6 to 14, this revision includes 36 of the original 85 items. The tests, which in this revision are arranged in order of difficulty, were chosen on the basis of age correlation, reliability, and certain practical considerations. Tentative percentile norms were found on a standardization sample of 380 boys and 369 girls attending public schools in central Illinois. Split-half reliabilities computed for single age and sex groups fell mostly in the .80's and .90's. A one-year retest yielded a correlation of .70. A factor analysis of a slightly longer, earlier version indicated a single common factor identified as motor development.

The Vineland Social Maturity Scale (Doll, 1953, 1965) is a developmental schedule concerned with the individual's ability to look after his practical needs and to take responsibility. Although covering a range from birth to over 25 years, this scale has been found most useful at the younger age levels, and particularly with the mentally retarded. The entire scale consists of 117 items grouped into year levels. The information required for each item is obtained, not through test situations, but through an interview with an informant or with the examinee himself. The scale is based on what the individual has actually done in his daily living. The items fall into eight categories: general self-help, self-help in eating, self-help in dressing, self-direction, occupation, communication, locomotion, and socialization. A social age (SA) and a social quotient (SQ) can be computed from the person's record on the entire scale.

The Vineland scale was standardized on 620 cases, including 10 males and 10 females at each year from birth to 30 years. These norms undoubtedly need updating. Moreover, the sample contained too few cases at each age and was not sufficiently representative of the general population, most of the cases coming from middle-class homes. A retest reliability of .92 is reported for 123 cases, the retest intervals varying from one day to nine months. The use of different examiners or informants did not appreciably affect results in this group, as long as all informants had had an adequate opportunity to observe the subjects.

---

[1] Other motor tests, designed specifically for use in the diagnosis of learning disabilities, will be considered in Chapter 16.

Validity of the scale was determined chiefly on the basis of age differentiation, comparison of normals with mental retardates, and correlation of scores with judgments of observers who knew the subjects well. Correlations between the Vineland scale and the Stanford-Binet vary widely, but are sufficiently low, in general, to indicate that different facets of behavior are being tapped by the two scales. The Vineland Social Maturity Scale has proved helpful to clinicians in diagnosing mental retardation and in reaching decisions regarding institutionalization. For example, an individual who is intellectually deficient in terms of the Stanford-Binet may be able to adjust satisfactorily outside an institution if his social age on the Vineland scale is adequate.

A newer and more comprehensive instrument is the Adaptive Behavior Scale (ABS), prepared by a committee of the American Association on Mental Deficiency. Designed primarily for mental retardates, this scale can also be used with emotionally maladjusted and other handicapped persons. Adaptive behavior is defined as "the effectiveness of an individual in coping with the natural and social demands of his or her environment" (American Association on Mental Deficiency, 1974). In its 1974 revision, this scale provides a single form applicable from the age of 3 years on. Like the Vineland, it is based on observations of everyday behavior and may be completed by parents, teachers, ward personnel, or others who have been in close contact with the examinee. The information may also be obtained through questioning or interviewing of one or more observers.

The ABS consists of two parts. Part 1 is a developmental scale covering 10 behavior domains, several of which are divided into subdomains as indicated below:

Independent Functioning: eating, toilet use, cleanliness, appearance, care of clothing, dressing and undressing, travel, general independent functioning

Physical Development: sensory development, motor development

Economic Activity: money handling and budgeting, shopping skills

Language Development: expression, comprehension, social language development

Numbers and Time

Domestic Activity: cleaning, kitchen duties, other domestic activities

Vocational Activity

Self-Direction: initiative, perseverance, leisure time

Responsibility

Socialization

Specific items with multiple responses are provided under each domain or subdomain. Special instructions are included for handling questions about activities that the individual may have had no opportunity to perform (e.g., shopping, restaurant eating).

Part 2 is designed to assess maladaptive behavior related to personality and behavior disorders. It covers 14 behavior domains, such as violent and destructive behavior, withdrawal, and hyperactive tendencies. In each of the 14 categories, specific items of behavior (e.g., "bites others," "attempts to set fires") are rated 1 if they occur occasionally and 2 if they occur frequently.

Instructions for administering and scoring are clearly set forth on the form itself and explained further in the manual. The scale yields a summary profile of percentile scores in each of the 24 behavior domains. Norms were obtained on institutionalized mental retardates of both sexes between the ages of 3 and 69 years. Percentile equivalents are reported for 11 age levels, grouped by one-year intervals at the youngest ages and by 2, 3, 10, and 20 years at the older ages. Cases were drawn from institutions throughout the United States, the numbers at each age level ranging from about 100 to slightly over 500.

In the interpretation of scores, the authors emphasize the importance of considering the individual's ability to remain and to function adequately in his own setting, community, or neighborhood. Preliminary data on rater reliability and on concurrent and construct validity are promising. The authors refer to several lines of research under way, including the application of the scale to noninstitutionalized mental retardates and to emotionally disturbed but nonretarded persons, longitudinal studies of change in scores during treatment and training programs, and investigations of various psychometric features of the instrument itself.

All the developmental scales discussed in this section fall short of psychometric requirements, particularly with reference to representativeness of normative data. Within the framework of current test-construction procedures, they all require more research to permit adequate interpretation of test results. Their principal interest stems from their indicating the scope of functions that need to be included in the examination of young children and of older children with mental or physical disorders.

BAYLEY SCALES OF INFANT DEVELOPMENT. The decades of the 1960s and the 1970s witnessed an upsurge of interest in tests for infants and preschool children. One contributing factor was the rapid expansion of educational programs for mentally retarded children. Another was the widespread development of preschool programs of compensatory education for culturally disadvantaged children. To meet these pressing practical needs, new tests have appeared and considerable research has been conducted on innovative approaches to assessment.[2]

An especially well-constructed test for the earliest age levels is the Bayley Scales of Infant Development, illustrated in Figure 37. Incorporating some items from the Gesell schedules and other infant and preschool tests, these scales represent the end-product of many years of research by Bayley and her co-workers, including the longitudinal investigations of the Berkeley Growth Study.

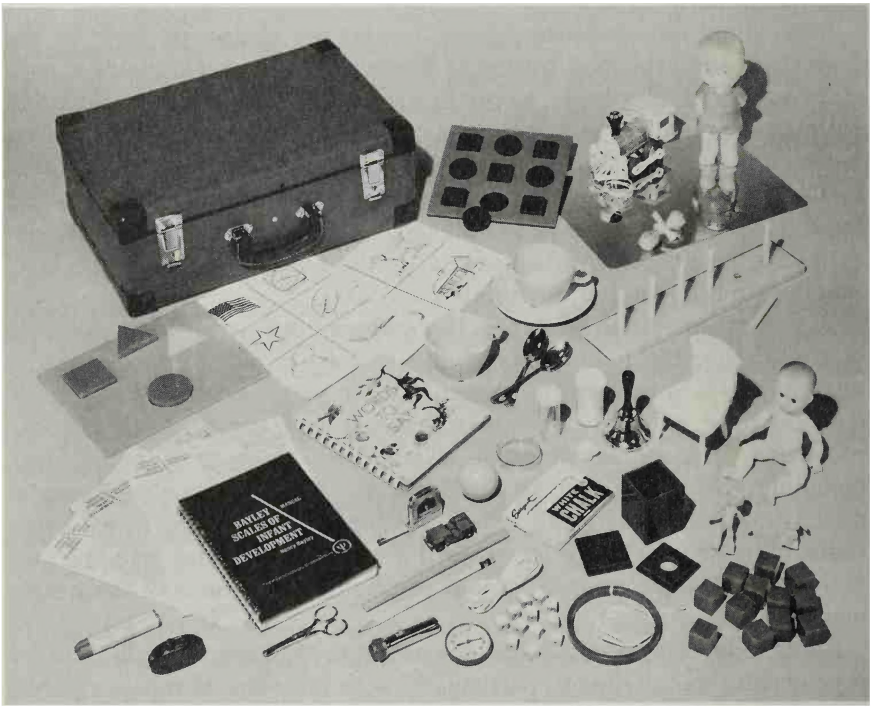


Fig. 37. Test Objects Employed with the Bayley Scales of Infant Development.

(Courtesy The Psychological Corporation.)

The Bayley scales provide three complementary tools for assessing the developmental status of children between the ages of 2 months and 2½ years: the Mental Scale, the Motor Scale, and the Infant Behavior Record.

---

[2] Surveys of available tests for infant and preschool levels can be found in Stott and Ball (1965) and Thomas (1970). The Wechsler Preschool and Primary Scale of Intelligence, discussed in Chapter 9, also belongs in the present category.

The Mental Scale samples such functions as perception, memory, learning, problem solving, vocalization, the beginnings of verbal communication, and rudimentary abstract thinking. The Motor Scale provides measures of gross motor abilities, such as sitting, standing, walking, and stair climbing, as well as manipulatory skills of hands and fingers. At the infant level, locomotor and manipulatory development plays an important part in the child's interactions with his environment and hence in the development of his mental processes. The Infant Behavior Record is a rating scale completed by the examiner after the other two parts have been administered. It is designed to assess various aspects of personality development, such as emotional and social behavior, attention span, persistence, and goal directedness.

In the technical quality of their test-construction procedures, the Bayley scales are clearly outstanding among tests for the infant level. Norms were established on 1,262 children, about equally distributed between the ages of 2 and 30 months. The standardization sample was chosen so as to be representative of the U.S. population in terms of urban-rural residence, major geographic region, sex ratio, race (white-nonwhite), and education of head of household. Institutionalized children, premature babies, and children over the age of 12 months from bilingual homes were excluded. Mental and Motor scales yield separate developmental indexes, expressed as normalized standard scores with a mean of 100 and an *SD* of 16 (as in Stanford-Binet deviation IQ's). These developmental indexes are found within the child's own age group, classified by half-month steps from 2 to 6 months and by one-month steps from 6 to 30 months.

Split-half reliability coefficients within separate age groups ranged from .81 to .93, with a median value of .88, for the Mental Scale; and from .68 to .92, with a median of .84, for the Motor Scale. These coefficients compare favorably with the reliabilities usually found in testing infants. The manual reports standard errors of measurement and minimum differences between indexes on the Mental and Motor scales required for statistical significance. Data on tester-observer agreement and on retest reliability after a one-week interval are also encouraging.

Bayley observes that these scales, like all infant tests, should be used principally to assess current developmental status rather than to predict subsequent ability levels. Development of abilities at these early ages is susceptible to so many intervening influences as to render long-term predictions of little value. The scales can be most helpful in the early detection of sensory and neurological defects, emotional disturbances, and environmental deficits.

MCCARTHY SCALES OF CHILDREN'S ABILITIES. At the preschool level, a recently developed instrument is the McCarthy Scales of Children's

Abilities (MSCA), suitable for children between the ages of 2½ and 8½ years. It consists of 18 tests, grouped into six overlapping scales: Verbal, Perceptual-Performance, Quantitative, General Cognitive, Memory, and Motor. Figure 38 illustrates the Conceptual Grouping Test from the Perceptual-Performance Scale. In this test, the child is shown red, yellow, and blue circles and squares in two sizes and asked to find the pieces with stated characteristics. The General Cognitive score comes closest to the traditional global measure of intellectual development. It is found from the sum of the scores on the first three scales, which do not overlap each other but which do contain all the Memory tests and all but three of the Motor tests. The General Cognitive score is thus based on 15 of the 18 tests in the entire battery.

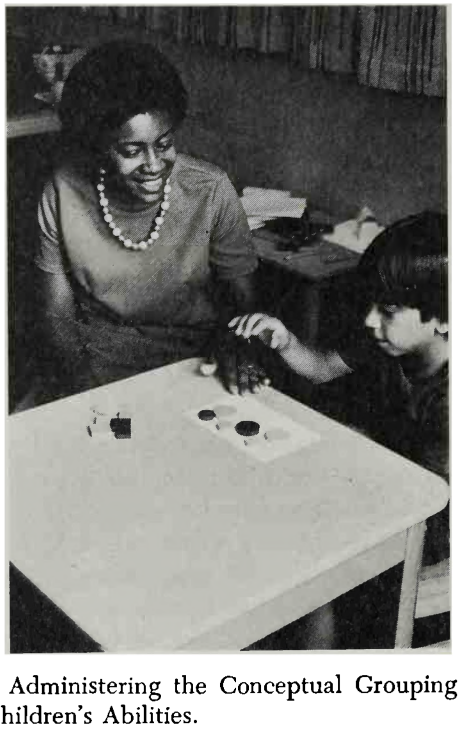


Fig. 38. Examiner Administering the Conceptual Grouping Test of the McCarthy Scales of Children's Abilities.

(Courtesy The Psychological Corporation.)

Scores on the General Cognitive Scale are expressed as a General Cognitive Index (GCI), which is a normalized standard score with a mean of 100 and an *SD* of 16, found within each 3-month age group. The manual makes it clear that, although the GCI is in the same units as traditional IQ's, the term IQ was deliberately avoided because of its many misleading connotations. The GCI is described as an index of the child's functioning at the time of testing, with no implications of immutability or etiology. Scores on the separate scales are normalized standard scores with a mean of 50 and an *SD* of 10 in terms of the same age groups.

The standardization sample of 1,032 cases included approximately 100 children at each of 10 age levels, by half-year steps between 2½ and 5½ and by one-year steps between 5½ and 8½. At each age level, the sample contained an equal number of boys and girls and was stratified by color (white-nonwhite), geographic region, father's occupational level, and (approximately) urban-rural residence, in accordance with the 1970 U.S. Census. Institutionalized mental retardates and children with severe behavioral or emotional disorders, brain damage, or obvious physical defects were excluded; bilinguals were tested only if they could speak and understand English.

Split-half reliability for the General Cognitive Index averaged .93 within age levels; average coefficients for the other five scales ranged from .79 to .88. The manual also reports standard errors of measurement and minimum differences between scale scores required for significance at the .05 level. Retest reliabilities over a one-month interval for 125 children classified into three age groups averaged .90 for GCI and ranged from .69 to .89 for the separate scales.

With regard to validity, the manual cites suggestive but meager data on predictive validity against an educational achievement battery administered at the end of the first grade. The initial selection of tests and their grouping into scales was based on a combination of clinical experience, findings of developmental psychology, and the results of factor-analytic research. In the course of developing the scales, a somewhat longer series of tests was factor analyzed separately within three age levels on about 60 percent of the standardization sample (Kaufman & Hollenbeck, 1973). General cognitive, memory, and motor factors were identified at all ages; other factors showed developmental trends, suggesting that different abilities might be utilized in performing the same tasks at different ages. For example, drawing tests had a large motor component at the younger ages but were predominantly conceptual at the older ages. The results of the preliminary factor analyses were substantially corroborated in subsequent factor analyses of the final version of the battery on the entire standardization sample (Kaufman, 1975b).

Other studies have investigated differences in MSCA performance in relation to sex, race (black-white) and socioeconomic status (father's occupational level). No significant sex differences were found in either GCI or any of the separate scale indexes (Kaufman & Kaufman, 1973b). Ethnic comparisons (Kaufman & Kaufman, 1973a) revealed few significant differences: in favor of blacks on the Motor Scale in the youngest age group (4–5½); and in favor of whites on the other scales in the oldest age group (6½–8½). Moreover, investigations of paternal occupational level within both ethnic groups suggest that socioeconomic status may be more important than race as a concomitant of performance on the McCarthy scales (Kaufman & Kaufman, 1975).

PIAGETIAN SCALES. Although applicable well beyond the preschool level, the scales modeled on the developmental theories of Jean Piaget have thus far found their major applications in early childhood. All such scales are in an experimental form; few are commercially available. Most have been developed for use in the authors' own research programs, although some of these scales are available to other research workers. At this stage, the major contribution of Piagetian scales to the psychological testing of children consists in their providing a theoretical framework that focuses on developmental sequences and a procedural approach characterized by flexibility and qualitative interpretation.

Some of the features of Piagetian scales, with special reference to normative interpretation of performance, were discussed in Chapter 4. Basically, Piagetian scales are ordinal in the sense that they presuppose a uniform sequence of development through successive stages. They are also content-referenced, insofar as they provide qualitative descriptions of what the child is actually able to do. Piagetian tasks focus on the long-term development of specific concepts or cognitive schemata,[3] rather than on broad traits. With regard to administration, the major object of Piagetian scales is to elicit the child's explanation for an observed event and the reasons that underlie his explanation. Scoring is characteristically based on the quality of the child's responses to a relatively small number of problem situations presented to him, rather than on the number or difficulty of successfully completed items. For this purpose, misconceptions revealed by the incorrect responses are of primary interest. The examiner concentrates more on the process of problem solving than on the product.

Because of its highly individualized procedures, Piagetian testing is well suited for clinical work. It has also attracted the attention of educators because it permits the integration of testing and teaching.[4] Its most frequent use, however, is still in research on developmental psychology. The three sets of tests described below have been selected in part because of their present or anticipated availability for use outside of the authors' own research, and in part because of their detailed description in published sources.[5]

---

[3] "Schemata" is the plural of "schema," a term commonly encountered in Piagetian writings and signifying essentially a framework into which the individual fits incoming sensory data.

[4] An example of such an application is "Let's Look at Children," to be discussed in Chapter 14.

At the University of Montreal, Laurendeau and Pinard have been engaged in an unusually comprehensive, long-term research project designed to replicate Piaget's work under standardized conditions, with large representative samples, and in a different cultural milieu (Laurendeau & Pinard, 1962, 1970; Pinard & Laurendeau, 1964). A byproduct of this research is the construction of scales of mental development that will eventually be available to other investigators. At this time, however, the authors feel it would be premature to release their scale of mental development until they have completed much more research with their tests.[6]

In the course of their investigations, Laurendeau and Pinard administered a battery of 57 tests to 700 children ranging in age from 2 to 12 years. The tests for children under 4 were newly constructed or adapted from conventional scales, although all were chosen to assess characteristics that Piaget attributed to this developmental period. Twenty-five of the tests, designed chiefly for ages 4 and up, were modeled directly after Piagetian tasks. Thus far, the results obtained with 10 of these tests have been reported in detail in two books (Laurendeau & Pinard, 1962, 1970). Five tests deal with *causality*, including the child's explanations of the nature and causes of dreams, the differences between animate and inanimate objects, what makes it dark at night, what causes the movements of clouds, and why some objects float and others sink. These tests are administered almost entirely through oral questionnaires, which fall about midway between Piaget's unstructured "methode clinique" and the completely controlled techniques of traditional tests. All questions are standardized; but depending upon the child's initial responses, the examiner follows alternative routes in his further exploration of the child's thinking processes.

The other five tests are concerned with the child's concepts of *space*. They include such tasks as recognizing objects by touch and identifying them among visually presented drawings of the same objects; arranging a set of toy lampposts in a straight line between two toy houses; placing a toy man in the same spots in the child's landscape that he occupies in the examiner's identical landscape; designating right and left on the child's own body, on the examiner in different positions, and in the relation of objects on the table; and problems of perspective, in which the child indicates how three toy mountains look to a man standing in different places. Several of these spatial tests deal with the "egocentrism" of the young child's thinking, which makes it difficult for him to regard objects from viewpoints other than his own.

---

[5] Other major projects on the standardization of Piagetian scales are being conducted by S. K. Escalona at the Albert Einstein College of Medicine in New York City, R. D. Tuddenham at the University of California in Berkeley, N. P. Vinh-Bang and B. Inhelder in Piaget's laboratory in Geneva, Switzerland, and E. A. Lunzer at the University of Manchester in England.

[6] Personal communication from Professor A. Pinard, April 3, 1974.

The complete protocol of the child's responses to each test is scored as a unit, in terms of the developmental level indicated by the quality of the responses. Laurendeau and Pinard have subjected their tests to extensive statistical analyses. Their standardization sample of 700 cases included 25 boys and 25 girls at each six-month interval from 2 to 5 years, and at each one-year interval from 5 to 12. The children were selected so as to constitute a representative sample of the French Canadian population of Montreal with regard to father's occupational level and school grade (or number of children in the family at the preschool ages). Besides providing normative age data, the authors analyzed their results for ordinality, or uniformity of sequence in the attainment of response levels by different children. They also investigated the degree of similarity in the developmental stages reached by each child in the different tests. Intercorrelations of scores on the five causality tests ranged from .59 to .78; on the five space tests, the correlations ranged from .37 to .67 (Laurendeau & Pinard, 1962, p. 236; 1970, p. 412).

The Ordinal Scales of Psychological Development prepared by Uzgiris and Hunt (1975) are designed for a much younger age level than are those of Laurendeau and Pinard, extending from the age of 2 weeks to 2 years. These ages cover approximately what Piaget characterizes as the sensorimotor period and within which he recognizes six stages. In order to increase the sensitivity of their instruments, Uzgiris and Hunt classify the responses into more than six levels, the number varying from 7 to 14 in the different scales. The series includes six scales, designated as follows:

1. *Object Permanence*—the child's emerging notion of independently existing objects is indicated by visual following of an object and searching for an object after it is hidden with increasing degrees of concealment.
2. *Development of Means* for achieving desired environmental ends—use of own hands in reaching for objects and of other means such as strings, stick, support, etc.
3. *Imitation*—including both gestural and vocal imitation.
4. *Operational Causality*—recognizing and adapting to objective causality, ranging from visual observation of one's own hands to eliciting desired behavior from a human agent and activating a mechanical toy.
5. *Object Relations in Space*—coordination of schemata of looking and listening in localizing objects in space; understanding such relations as container, equilibrium, gravity.

6. *Development of Schemata* for relating to objects—responding to objects by looking, feeling, manipulating, dropping, throwing, etc., and by socially instigated schemata appropriate to particular objects (e.g., "driving" toy car, building with blocks, wearing beads, naming objects).

No norms are provided, but the authors collected data on several psychometric properties of their scales by administering them to 84 infants, including at least four at each month of age up to one year and at least four at each two months of age between one and two years. Most of these subjects were children of graduate students and staff at the University of Illinois. Both observer agreement and test-retest agreement after a 48-hour interval are reported. In general, the tests appear quite satisfactory in both respects. An index of ordinality, computed for each scale from the scores of the same 84 children, ranged from .802 to .991. The authors report that .50 is considered minimally satisfactory evidence of ordinality with the index employed.[7]

Uzgiris and Hunt clearly explain that these are only provisional scales, although they are available to other investigators for research purposes. Apart from journal articles reporting specific studies in which the scales were employed, the authors describe the tests in a book (Uzgiris & Hunt, 1975) and also provide six sound films demonstrating their use. The scales were originally designed to measure the effects of specific environmental conditions on the rate and course of development of infants. Studies of infants reared under different conditions (Paraskevopoulos & Hunt, 1971) and on infants participating in intervention programs (Hunt, Paraskevopoulos, Schickedanz, & Uzgiris, 1975) have thus far indicated significant effects of these environmental variables on the mean age at which children attain different steps in the developmental scales.

Unlike the first two examples of Piagetian scales, the Concept Assessment Kit—Conservation is a published test which may be purchased on the same basis as other psychological tests. Designed for ages 4 to 7 years, it provides a measure of one of the best known Piagetian concepts. Conservation refers to the child's realization that such properties of objects as weight, volume, or number remain unchanged when the objects undergo transformations in shape, position, form, or other specific attributes. The authors (Goldschmid & Bentler, 1968b) focused on conservation as an indicator of the child's transition from the preoperational to the concrete operational stage of thinking, which Piaget places roughly at the age of 7 or 8 years.

Throughout the test, the procedure is essentially the same. The child is shown two identical objects; then the examiner makes certain transformations in one of them and interrogates the child about their similarity or difference. After answering, the child is asked to explain his answer. In each item, one point is scored for the correct judgment of equivalence and one point for an acceptable explanation. For example, the examiner begins with two standard glasses containing equal amounts of water (continuous quantity) or grains of corn (discontinuous quantity) and pours the contents into a flat dish or into several small glasses. In another task, the examiner shows the child two equal balls of Playdoh and then flattens one into a pancake and asks whether the ball is as heavy as the pancake.

---

[7] Procedures for the measurement of ordinality and the application of scalogram analysis to Piagetian scales are still controversial, a fact that should be borne in mind in interpreting any reported indices of ordinality (see Hooper, 1973; Wohlwill, 1970).

Three forms of the test are available. Forms A and B are parallel, each providing six tasks: Two-Dimensional Space, Number, Substance, Continuous Quantity, Weight, and Discontinuous Quantity. The two forms were shown to be closely equivalent in means and *SD*'s and their scores correlated .95. Form C includes two different tasks, Area and Length; it correlates .76 and .74 with Forms A and B, respectively. Administration is facilitated by printing all essential directions on the record form, including diagrams of the materials, directions for manipulating materials, and verbal instructions.

Norms were established on a standardization sample of 560 boys and girls between the ages of 4 and 8, obtained from schools, day care centers, and Head Start centers in the Los Angeles, California area. The sample included both blacks and whites and covered a wide range of socioeconomic level, but with a slight overrepresentation of the lower-middle class. Percentile norms are reported for each age level. These norms, of course, must be regarded as tentative in view of the small number of cases at each age and the limitations in representativeness of the sample. Mean scores for each age show a systematic rise with age, with a sharp rise between 6 and 8 years, as anticipated from Piagetian theory.

Both in the process of test construction and in the evaluation of the final forms, the authors carried out various statistical analyses to assess scorer reliability; Kuder-Richardson, parallel-form, and retest reliability; scalability, or ordinality; and factorial composition (see also Goldschmid & Bentler, 1968a). Although based on rather small samples, the results indicate generally satisfactory reliability and give good evidence of ordinality and of the presence of a large common factor of conservation throughout the tasks.

Comparative studies in seven countries suggest that the test is applicable in widely diverse cultures, yielding high reliabilities and showing approximately similar age trends (Goldschmid et al., 1973). Differences among cultures and subcultures, however, have been found in the mean ages at which concepts are acquired, i.e., the age curves may be displaced horizontally by one or two years (see also Figurelli & Keller, 1972; Wasik

& Wasik, 1971). Training in conservation tasks has been found to improve scores significantly (see also Goldschmid, 1968; Zimmerman & Rosenthal, 1974a, 1974b). The manual cites several studies on small groups that contribute suggestive data about the construct validity of the test. Some evidence of predictive validity is provided by significant correlations in the .30's and .40's with first-grade achievement, the correlation being highest with arithmetic grades (.52).

TESTING THE PHYSICALLY HANDICAPPED

DEAFNESS. Owing to their general retardation in linguistic development, deaf children are usually handicapped on verbal tests, even when the verbal content is presented visually. In fact, the testing of deaf children was the primary object in the development of some of the earliest performance scales, such as the Pintner-Paterson Performance Scale and the Arthur Performance Scale. In Revised Form II of the Arthur scale, the verbal instructions required in the earlier form were further reduced in order to increase the applicability of the test to deaf children. Special adaptations of the Wechsler scales are sometimes employed in testing deaf persons. The verbal tests can be administered if the oral questions are typed on cards. Various procedures for communicating the instructions for the performance tests have also been worked out (see Sattler, 1974, pp. 170–172). With such modifications of standard testing procedures, however, one cannot assume that reliability, validity, and norms remain unchanged. Nonlanguage group tests, such as the Army Beta, are also used in testing the deaf.

Whether or not they require special procedural adaptations, all the tests mentioned thus far were standardized on hearing persons. For many purposes, it is of course desirable to compare the performance of the deaf with general norms established on hearing persons. At the same time, norms obtained on deaf children are also useful in a number of situations pertaining to the educational development of these children.

To meet this need, the Hiskey-Nebraska Test of Learning Aptitude was developed and standardized on deaf and hard-of-hearing children. This is an individual test suitable for ages 3 to 16. Speed was eliminated, since it is difficult to convey the idea of speed to young deaf children. An attempt was also made to sample a wider variety of intellectual functions than those covered by most performance tests. Pantomime and practice exercises to communicate the instructions, as well as intrinsically interesting items to establish rapport, were considered important requirements for such a test. All items were chosen with special reference to the limitations of deaf children, the final item selection being based chiefly on the criterion of age differentiation.

The Hiskey-Nebraska Test consists of twelve subtests:

1. Bead Patterns
2. Memory for Color
3. Picture Identification
4. Picture Associations
5. Paper Folding (Patterns)
6. Visual Attention Span
7. Block Patterns
8. Completion of Drawings
9. Memory for Digits
10. Puzzle Blocks
11. Picture Analogies
12. Spatial Reasoning

Norms were derived separately from 1,079 deaf and 1,074 hearing children between the ages of 3 and 17 years, tested in 10 states. Split-half reliabilities in the .90's are reported for deaf and hearing groups. Intercorrelations of the 12 subtests range from the .30's to the .70's among younger children (ages 3 to 10) and from the .20's to the .40's among older children (ages 11 to 17). Correlations of .78 to .86 were found between the Hiskey-Nebraska and either the Stanford-Binet or the Wechsler Intelligence Scale for Children in small groups of hearing children. Further evidence of validity was provided by substantial correlations with achievement tests among deaf children. The manual contains a discussion of desirable practices to be followed in testing deaf children.

BLINDNESS. Testing the blind presents a very different set of problems from those encountered with the deaf. Oral tests can be most readily adapted for blind persons, while performance tests are least likely to be applicable. In addition to the usual oral presentation by the examiner, other suitable testing techniques have been utilized, such as phonograph records and tape or wire recordings. Some tests are also available in braille. The latter technique is somewhat limited in its applicability, however, by the greater bulkiness of materials printed in braille as compared with inkprint, by the slower reading rate for braille, and by the number of blind persons who are not facile braille readers. The examinee's responses may likewise be recorded in braille or on a typewriter. Specially prepared embossed answer sheets or cards are also available for use with true-false, multiple-choice, and other objective-type items. In many individually administered tests, of course, oral responses can be obtained.

Among the principal examples of general intelligence tests that have been adapted for blind persons are the Binet and the Wechsler. The first Hayes-Binet revision for testing the blind was based on the 1916 Stanford-Binet. In 1942, the Interim Hayes-Binet[8] was prepared from the 1937

---

[8] Originally designated as an interim edition because of the tentative nature of its standardization, this revision has come to be known by this name in the literature.

Stanford-Binet (S. P. Hayes, 1942, 1943). All items that could be administered without the use of vision were selected from both Form L and Form M. This procedure yielded six tests for each year level from VIII to XIV, and eight tests at the Average Adult level. In order to assemble enough tests for year levels III to VI, it was necessary to draw on some of the special tests devised for use in the earlier Hayes-Binet. Most of the tests in the final scale are oral, a few requiring braille materials. A retest reliability of .90 and a split-half reliability of .91 are reported by Hayes. Correlations with braille editions of standard achievement tests ranged from .82 to .93. The validity of this test was also checked against school progress.

The Wechsler scales have also been adapted for blind examinees. These adaptations consist essentially in using the verbal tests and omitting the performance tests. A few items inappropriate for the blind have also been replaced by alternates. When tested under these conditions, blind persons as a group have been found to equal or excel the general seeing norms.

A different approach is illustrated by the Haptic Intelligence Scale for Adult Blind. This was developed as a nonverbal test to be used in conjunction with the Verbal Scale of the WAIS. Four of the tests are adaptations of performance tests from the WAIS, namely, Digit Symbol, Block Design, Object Assembly, and Object Completion; two were newly devised, including Pattern Board and Bead Arithmetic. The tests utilize a completely tactile approach and, if given to the partially sighted, require the wearing of a blindfold. For this reason, among others, they are probably best suited for testing the totally blind. Standardization procedures followed closely those employed with the WAIS. The blind subjects tested in the standardization sample included a proportional number of nonwhites and were distributed over the major geographical regions of the country. Subtest scores and deviation IQ's are found as in the WAIS. Split-half reliability for the entire test was found to be .95 and for subtests varied from .79 (Object Assembly) to .94 (Bead Arithmetic). A six-month retest of 136 subjects yielded a total-score reliability of .91 and subtest reliabilities ranging from .70 to .81. Correlation with WAIS Verbal Scale in the 20–34 age group of blind subjects was .65. The materials are bulky and administration time is long, requiring from 1½ to 2 hours; but blind examinees generally find the tests interesting and enjoyable. The authors caution that this is a provisional scale, requiring further research. It can provide useful information when employed by a trained clinician.

A number of group intelligence tests have likewise been adapted for use with the visually handicapped and are available in both large-type and braille editions. Examples include the School and College Ability Tests (SCAT), the College Board Scholastic Aptitude Test (SAT), and the Aptitude Test of the Graduate Record Examinations (GRE). Research with a tactile form of the Progressive Matrices has shown it to have promise as a nonverbal intelligence test for blind children between the ages of 9 and 15 years (Rich & Anderson, 1965). An adaptation of the Vineland Social Maturity Scale for blind preschool children was developed and standardized by Maxfield and Buchholz (1957).

ORTHOPEDIC HANDICAPS. Although usually able to receive auditory and visual stimulation, the orthopedically handicapped may have such severe motor disorders as to make either oral or written responses impracticable. The manipulation of formboards or other performance materials would likewise meet with difficulties. Working against a time limit or in strange surroundings often increases the motor disturbance in the orthopedically handicapped. Their greater susceptibility to fatigue makes short testing sessions necessary.

Some of the severest motor handicaps are found among the cerebral palsied. Yet surveys of these cases have frequently employed common intelligence tests such as the Stanford-Binet or the Arthur Performance Scale. In such studies, the most severely handicapped were usually excluded as untestable. Frequently, informal adjustments in testing procedure are made in order to adapt the test to the child's response capacities. Both of these procedures, of course, are makeshifts.

A more satisfactory approach lies in the development of testing instruments suitable for even the most severely handicapped individuals. A number of specially designed tests or adaptations of existing tests are now available for this purpose, although their normative and validity data are usually meager. Several of the tests to be discussed in the next section, originally designed for use in cross-cultural testing, have also proved applicable to the handicapped. Adaptations of the Leiter International Performance Scale and the Porteus Mazes, suitable for administration to cerebral-palsied children, have been prepared (Allen & Collins, 1955; Arnold, 1951). In both adapted tests, the examiner manipulates the test materials, while the subject responds only by appropriate head movements. A similar adaptation of the Stanford-Binet has been proposed (E. Katz, 1958). The Progressive Matrices provide a promising tool for this purpose. Since this test is given with no time limit, and since the response may be indicated orally, in writing, or by pointing or nodding, it appears to be especially appropriate for the orthopedically handicapped. Despite the flexibility and simplicity of its response indicator, this test covers a wide range of difficulty and provides a fairly high test ceiling. Successful use of this test has been reported in studies of cerebral-palsied children and adults (Allen & Collins, 1955; Holden, 1951; Tracht, 1948).

Another type of test that permits the utilization of a simple pointing response is the *picture vocabulary test.* These tests provide a rapid measure of "use" vocabulary, especially applicable to persons unable to vocalize well (such as the cerebral palsied) and to the deaf. Since they are easy to administer and can be completed in about 15 minutes or less, they are also useful as a rapid screening device in situations where no trained examiner is available but individual testing is needed.

The Peabody Picture Vocabulary Test (PPVT) is typical of these instruments. It consists of a series of 150 plates, each containing four pictures. As each plate is presented, the examiner provides a stimulus word orally; the child responds by pointing to or in some other way designating the picture on the plate that best illustrates the meaning of the stimulus word. Although the entire test covers a range from 2½ to 18 years, each individual is given only the plates appropriate to his own performance level, as determined by a specified run of successes at one end and failures at the other. Raw scores can be converted to mental ages, deviation IQ's, or percentiles. The PPVT is untimed but requires from 10 to 15 minutes. It is available in two parallel forms which utilize the same set of cards with different stimulus words.

The standardization sample for the PPVT included a total of 4,012 cases between the ages of 2½ and 18 years tested in Nashville, Tennessee, and its environs. Alternate form reliability coefficients for different age levels within the standardization sample ranged from .67 to .84. Reliability coefficients within the same range were subsequently obtained in several mentally retarded or physically handicapped groups. Validity was originally established in terms of age differentiation. Since its publication, the test has been employed in a number of studies with normal, mentally retarded, emotionally disturbed, or physically handicapped children. These studies have yielded validity coefficients in the .60's with individual and group intelligence scales within relatively homogeneous age groups. Understandably, these correlations were higher with verbal than with performance tests. There is also some evidence of moderate concurrent and predictive validity against academic achievement tests. A limitation of this test for certain testing purposes is suggested by the finding that culturally disadvantaged children tend to perform more poorly on it than on other intelligence tests (Costello & Ali, 1971; Cundick, 1970; Milgram & Ozer, 1967; Rosenberg & Stroud, 1966). On the other hand, participants in preschool compensatory education programs showed more improvement on this test than on the Stanford-Binet (Howard & Plant, 1967; Klaus & Gray, 1968; Milgram, 1971). Scores on the PPVT may reflect in part the child's degree of cultural assimilation.

Similar procedures of test administration have been incorporated in *pictorial classification tests*, as illustrated by the Columbia Mental Maturity Scale (CMMS). Originally developed for use with cerebral-palsied children, this scale comprises 92 items, each consisting of a set of three, four, or five drawings printed on a large card. The examinee is required to identify the drawing that does not belong with the others, indicating his choice by pointing or nodding (see Fig. 39). To heighten interest and appeal, the cards and drawings are varicolored. The objects depicted were chosen to be within the range of experience of most American children. Scores are expressed as Age Deviation Scores, which are normalized standard scores within age groups, with a mean of 100 and an *SD* of 16. Percentile and stanine equivalents for these scores are also provided. To meet the demand for developmental norms, the manual includes a Maturity Index, indicating the age group in the standardization sample whose test performance is most similar to that of the child.

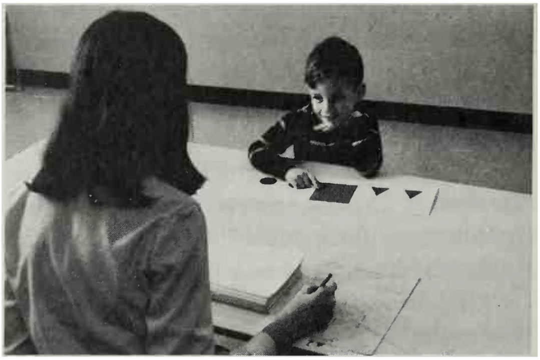


FIG. 39. Examiner Administering Columbia Mental Maturity Scale to Child. (From Columbia Mental Maturity Scale: Guide for Administering and Interpreting, 1972, p. 11. Copyright © 1972 by Harcourt Brace Jovanovich, Inc. Reproduced by permission.)

The standardization sample for the CMMS comprised 2,600 children, including 100 boys and 100 girls in each of 13 six-month age groups between the ages of 3–6 and 9–11. The sample was stratified in terms of the 1960 U.S. Census with regard to parental occupational level, race, and geographical region; proportion of children living in metropolitan and nonmetropolitan areas was also approximately controlled. Split-half reliabilities within single age groups ranged from .85 to .91. Standard errors of measurement of the Age Deviation Scores are between 5 and 6 points. Retest of three age groups after an interval of 7 to 10 days yielded reliabilities of .84 to .86. A correlation of .67 with Stanford-Binet was found in a group of 52 preschool and first-grade children. Correlations with achievement test scores in first- and second-grade samples fell mostly between the high .40's and the low .60's. More extensive data on validity and on applicability to various handicapped groups are available for an earlier form of the test.

CROSS-CULTURAL TESTING

THE PROBLEM. The testing of persons with highly dissimilar cultural backgrounds has received increasing attention since midcentury. Tests are needed for the maximum utilization of human resources in the newly developing nations in Africa and elsewhere. The rapidly expanding educational facilities in these countries require testing for admission purposes as well as for individual counseling. With increasing industrialization, there is a mounting demand for tests to aid in the job selection and placement of personnel, particularly in mechanical, clerical, and professional fields.

In America the practical problems of cross-cultural testing have been associated chiefly with subcultures or minority cultures within the dominant culture. There has been widespread concern regarding the applicability of available tests to culturally disadvantaged groups. It should be noted parenthetically that cultural disadvantage is a relative concept. Objectively there is only cultural difference between any two cultures or subcultures. Each culture fosters and encourages the development of behavior that is adapted to its values and demands. When an individual must adjust to and compete within a culture or subculture other than that in which he was reared, then cultural difference is likely to become cultural disadvantage.

Although concern with cross-cultural testing has been greatly stimulated by recent social and political developments, the problem was recognized at least as early as 1910. Some of the earliest cross-cultural tests were developed for testing the large waves of immigrants coming to the United States at the turn of the century. Other early tests originated in basic research on the comparative abilities of relatively isolated cultural groups. These cultures were often quite primitive and had had little or no contact with Western civilization within whose framework most psychological tests had been developed.

Traditionally, cross-cultural tests have tried to rule out one or more parameters along which cultures vary. A well-known example of such a parameter is *language*. If the cultural groups to be tested spoke different languages, tests were developed that required no language on the part of either examiner or subjects. When educational backgrounds differed widely and illiteracy was prevalent, *reading* was ruled out. Oral language was not eliminated from these tests because they were designed for persons speaking a common language. Another parameter in which cultures or subcultures differ is that of *speed*. Not only the tempo of daily life, but also the motivation to hurry and the value attached to rapid performance vary widely among national cultures, among ethnic minority groups within a single nation, and between urban and rural subcultures (see, e.g., Klineberg, 1928; Knapp, 1960). Accordingly, cross-cultural tests have often—though not always—tried to eliminate the influence of speed by allowing long time limits and giving no premium for faster performance.

Still other parameters along which cultures differ pertain to *test content.* Most nonlanguage and nonreading tests, for example, call for items of information that are specific to certain cultures. Thus, they may require the examinee to understand the function of such objects as violin, postage stamp, gun, pocketknife, telephone, piano, or mirror. Persons reared in certain cultures may lack the experiential background to respond correctly to such items. It was chiefly to control this type of cultural parameter that the classic "culture-free" tests were first developed. Following a brief examination of typical tests designed to eliminate one or more of the above parameters, we shall turn to an analysis of alternative approaches to cross-cultural testing.

TYPICAL INSTRUMENTS. In their efforts to construct tests applicable across cultures, psychometricians have followed a variety of procedures, some of which are illustrated by the four tests to be considered in this section. The Leiter International Performance Scale is an individually administered performance scale. It was developed through several years of use with different ethnic groups in Hawaii, including elementary and high school pupils. The scale was subsequently applied to several African groups by Porteus and to a few other national groups by other investigators. A later revision, issued in 1948, was based on further testing of American children, high school students, and Army recruits during World War II. A distinctive feature of the Leiter scale is the almost complete elimination of instructions, either spoken or pantomime. Each test begins with a very easy task of the type to be encountered throughout that test. The comprehension of the task is treated as part of the test. The materials consist of a response frame, illustrated in Figure 40, with an adjustable card holder. All tests are administered by attaching the appropriate card, containing printed pictures, to the frame. The examinee chooses the blocks with the proper response pictures and inserts them into the frame.

The Leiter scale was designed to cover a wide range of functions, similar to those found in verbal scales. Among the tasks included may be mentioned: matching identical colors, shades of gray, forms, or pictures; copying a block design; picture completion; number estimation; analogies; series completion; recognition of age differences; spatial relations; footprint recognition; similarities; memory for a series; and classification of animals according to habitat. Administered individually, with no time limit, these tests are arranged into year levels from 2 to 18. The scale is scored in terms of MA and ratio IQ, although there is no assurance that

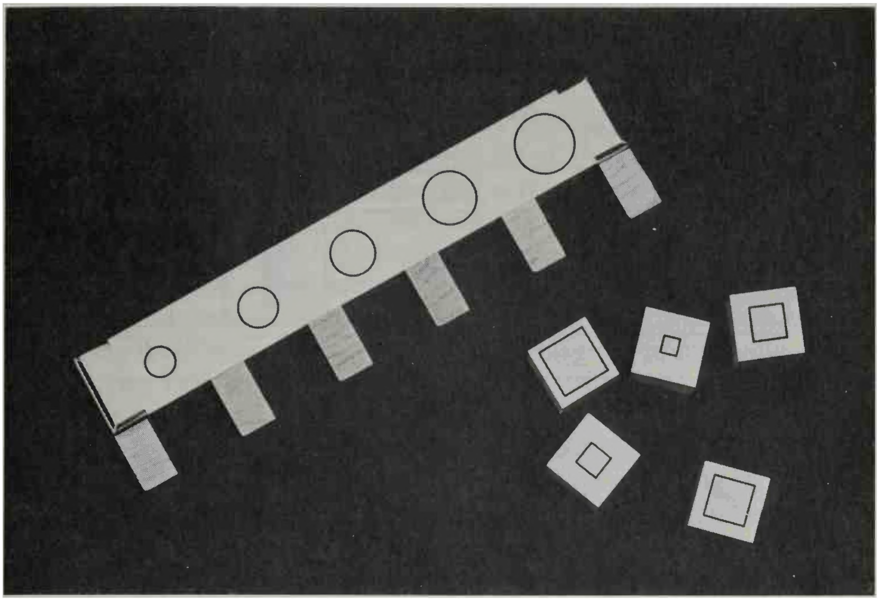


FIG. 40. Typical Materials for Use in the Leiter International Performance Scale. The test illustrated is the Analogies Progression Test from the Six-Year Level.

(Courtesy C. H. Stoelting Company.)

such an IQ retains the same meaning at different ages. In fact, the published data show considerable fluctuation in the standard deviation of the IQ's at different age levels. Split-half reliabilities of .91 to .94 are reported from several studies, but the samples were quite heterogeneous in age and probably in other characteristics. Validation data are based principally on age differentiation and internal consistency. Some correlations are also reported with teachers' ratings of intelligence and with scores on other tests, including the Stanford-Binet and the WISC. These correlations range from .56 to .92, but most were obtained on rather heterogeneous groups.

The tests in year levels 2 to 12 are also available as the Arthur Adaptation of the Leiter International Performance Scale. This adaptation, considered most suitable for testing children between the ages of 3 and 8 years, was standardized by Grace Arthur in 1952. Its norms must be regarded as quite limited, being derived from a standardization sample of 289 children from a middle-class, midwestern metropolitan background. Like the original scale, the Arthur adaptation yields an MA and a ratio IQ.

The Culture Fair Intelligence Test, developed by R. B. Cattell and published by the Institute for Personality and Ability Testing (IPAT), is a

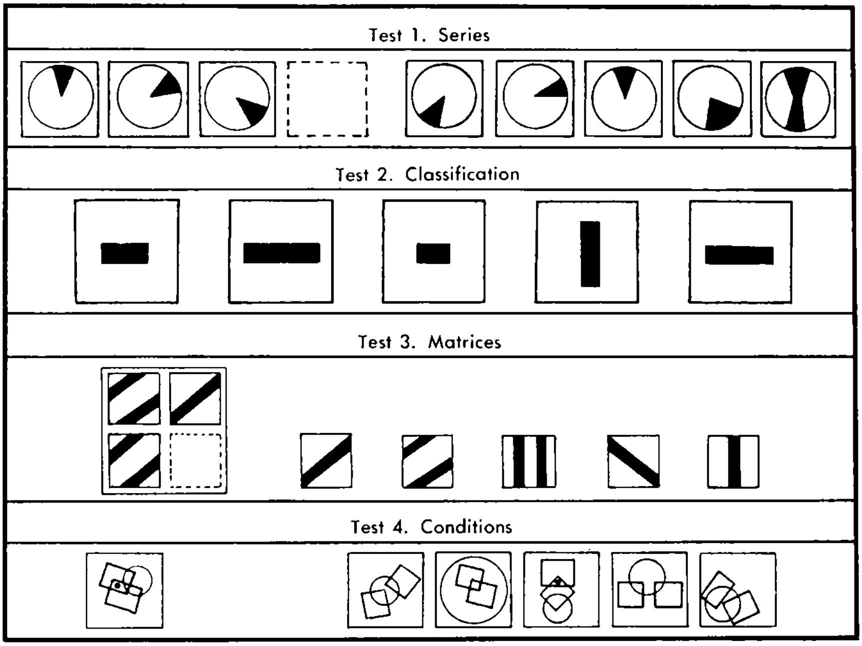


FIG. 41. Sample Items from Culture Fair Intelligence Test, Scale 2. (Copyright by Institute of Personality and Ability Testing.)

paper-and-pencil test. This test is available in three levels: Scale 1, for ages 4 to 8 and mentally retarded adults; Scale 2, for ages 8 to 13 and average adults; and Scale 3, for grades 10 to 16 and superior adults. Each scale has been prepared in two parallel forms, A and B. Scale 1 requires individual administration for at least some of the tests; the other scales may be given either as individual or as group tests. Scale 1 comprises eight tests, only four of which are described by the author as culture-fair. The other four involve both verbal comprehension and specific cultural information. It is suggested that the four culture-fair tests can be used as a sub-battery, separate norms being provided for this abbreviated scale. Scales 2 and 3 are alike, except for difficulty level. Each consists of the following four tests, sample items from which are shown in Figure 41.

1. *Series:* Select the item that completes the series.
2. *Classification:* Mark the one item in each row that does not belong with the others.
3. *Matrices:* Mark the item that correctly completes the given matrix, or pattern.
4. *Conditions:* Insert a dot in one of the alternative designs so as to meet the same conditions indicated in the sample design. Thus, in the example reproduced in Figure 41, the dot must be in the two rectangles, but not

in the circle. This condition can be met only in the third response alternative, which has been marked.

For Scale 1, only ratio IQ's are provided. In Scales 2 and 3, scores can be converted into deviation IQ's with an *SD* of 16 points. Scales 2 and 3 have been standardized on larger samples than Scale 1, but the representativeness of the samples and the number of cases at some age levels still fall short of desirable test-construction standards. Although the tests are highly speeded, some norms are provided for an untimed version. Fairly extensive verbal instructions are required, but the author asserts that giving these instructions in a foreign language or in pantomime will not affect the difficulty of the test.

Internal consistency and alternate-form reliability coefficients are marginal, especially for Scale 3, where they fall mostly in the .50's and .60's. Validity is discussed chiefly in terms of saturation with a general intellective factor (*g*), having been investigated largely through correlation with other tests and through factor analysis. Scattered studies of concurrent and predictive validity show moderate correlations with various academic and occupational criteria. The Cattell tests have been administered in several European countries, in America, and in certain African and Asian cultures. Norms tended to remain unchanged in cultures moderately similar to that in which the tests were developed; in other cultures, however, performance fell considerably below the original norms. Moreover, black children of low socioeconomic level tested in the United States did no better on this test than on the Stanford-Binet (Willard, 1968).

The Progressive Matrices, developed in Great Britain by Raven, were also designed as a measure of Spearman's *g* factor. Requiring chiefly the eduction of relations among abstract items, this test is regarded by most British psychologists as the best available measure of *g*. It consists of 60 matrices, or designs, from each of which a part has been removed. The subject chooses the missing insert from six or eight given alternatives. The items are grouped into five series, each containing 12 matrices of increasing difficulty but similar in principle. The earlier series require accuracy of discrimination; the later, more difficult series involve analogies, permutation and alternation of pattern, and other logical relations. Two sample items are reproduced in Figure 42. The test is administered with no time limit, and can be given individually or in groups. Very simple oral instructions are required.

Percentile norms are provided for each half-year interval between 8 and 14 years, and for each five-year interval between 20 and 65 years. These norms are based on British samples, including 1,407 children, 3,665 men in military service tested during World War II, and 2,192 civilian adults. Closely similar norms were obtained by Rimoldi (1948) on 1,680

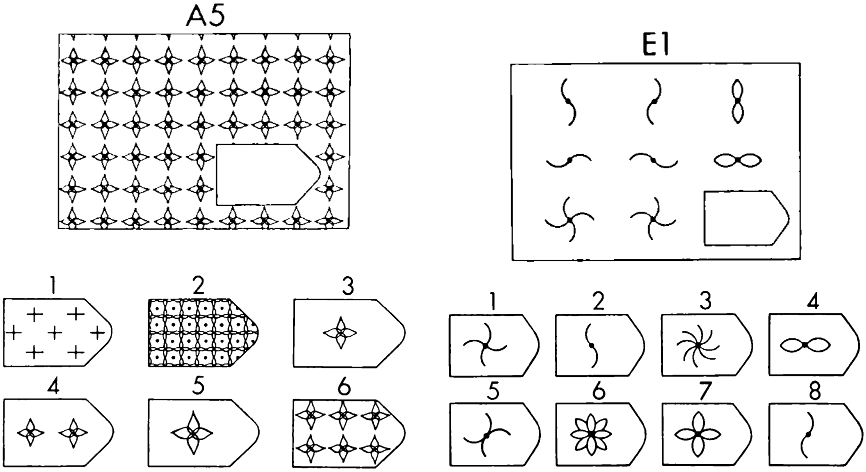


FIG. 42. Sample Items from the Progressive Matrices.
(Reproduced by permission of J. C. Raven.)

children in Argentina. Use of the test in several European countries likewise indicated the applicability of available norms. Studies in a number of non-European cultures, however, have raised doubts about the suitability of this test for groups with very dissimilar backgrounds. In such groups, moreover, the test was found to reflect amount of education and to be susceptible to considerable practice effect.

The manual for the Progressive Matrices is quite inadequate, giving little information on reliability and none on validity. Many investigations have been published, however, that provide relevant data on this test. In a review of publications appearing prior to 1957, Burke (1958) lists over 50 studies appearing in England, 14 in America, and 10 elsewhere. Since that time, research has continued at a rapid pace, especially in America where this test has received growing recognition. The *Seventh Mental Measurements Yearbook* lists nearly 400 studies, many dealing with the use of this test with clinical patients.

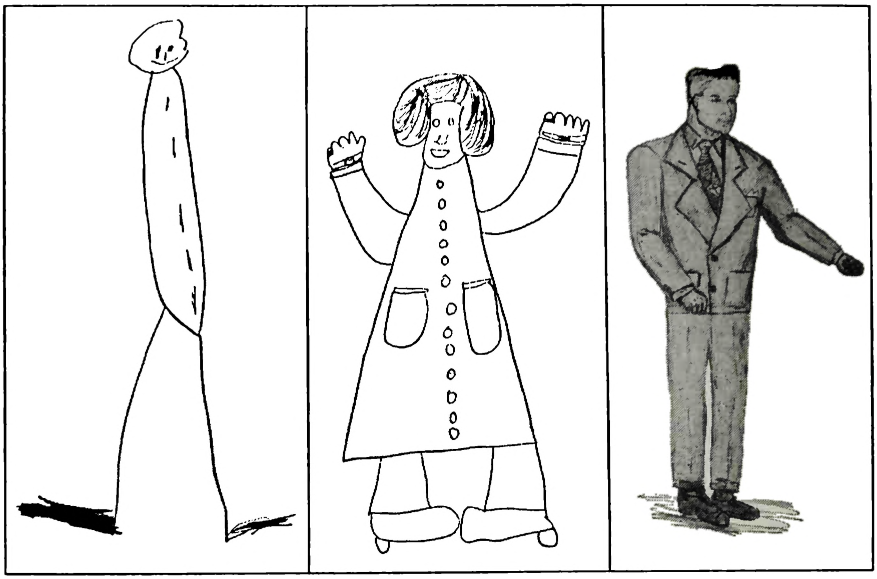
Retest reliability in groups of older children and adults that were moderately homogeneous in age varies approximately between .70 and .90. At the lower score ranges, however, reliability falls considerably below these values. Correlations with both verbal and performance tests of intelligence range between .40 and .75, tending to be higher with performance than with verbal tests. Studies with the mentally retarded and with different occupational and educational groups indicate fair concurrent validity. Predictive validity coefficients against academic criteria run somewhat lower than those of the usual verbal intelligence tests. Several factorial analyses suggest that the Progressive Matrices are heavily loaded with a factor common to most intelligence tests (identified with Spearman's *g* by British psychologists), but that spatial aptitude, inductive reasoning, perceptual accuracy, and other group factors also influence performance (Burke, 1958).

An easier form, the Coloured Progressive Matrices, is available for children between the ages of 5 and 11 years and for mentally retarded adults. A more advanced form has also been developed for superior adults, but its distribution is restricted to approved and registered users.

A still different approach is illustrated by the Goodenough Draw-a-Man Test, in which the examinee is simply instructed to "make a picture of a man; make the very best picture that you can." This test was in use without change from its original standardization in 1926 until 1963. An extension and revision was published in 1963 under the title of Goodenough-Harris Drawing Test (D. B. Harris, 1963). In the revision, as in the original test, emphasis is placed on the child's accuracy of observation and on the development of conceptual thinking, rather than on artistic skill. Credit is given for the inclusion of individual body parts, clothing details, proportion, perspective, and similar features. A total of 73 scorable items were selected on the basis of age differentiation, relation to total scores on the test, and relation to group intelligence test scores. Data for this purpose were obtained by testing samples of 50 boys and 50 girls at each grade level from kindergarten through the ninth grade in urban and rural areas of Minnesota and Wisconsin, stratified according to father's occupation.

In the revised scale, subjects are also asked to draw a picture of a woman and of themselves. The Woman scale is scored in terms of 71 items similar to those in the Man scale. The Self scale was developed as a projective test of personality, although available findings from this application are not promising. Norms on both Man and Woman scales were established on new samples of 300 children at each year of age from 5 to 15, selected so as to be representative of the United States population with regard to father's occupation and geographical region. Point scores on each scale are transmuted into standard scores with a mean of 100 and an *SD* of 15. In Figure 43 will be found three illustrative drawings produced by children aged 5–8, 8–8, and 12–11, together with the corresponding raw point scores and standard scores. An alternative, simplified scoring procedure is provided by the Quality scales for both Man and Woman drawings. Instead of the point scoring, the Quality scales utilize a global, qualitative assessment of the entire drawing, obtained by matching the child's drawing with the one it resembles most closely in a graded series of 12 samples.

The reliability of the Draw-a-Man Test has been repeatedly investigated by a variety of procedures. In one carefully controlled study of the earlier form administered to 386 third- and fourth-grade schoolchildren,



Man: Raw Score 7
CA 5-8
Standard Score 73

Woman: Raw Score 31
CA 8-8
Standard Score 103

Man: Raw Score 66
CA 12-11
Standard Score 134

FIG. 43. Specimen Drawings Obtained in Goodenough-Harris Drawing Test. (Courtesy Dale B. Harris.)

the retest correlation after a one-week interval was .68, and split-half reliability was .89 (McCarthy, 1944). Rescoring of the identical drawings by a different scorer yielded a scorer reliability of .90, and rescorings by the same scorer correlated .94. Studies with the new form (Dunn, 1967; D. B. Harris, 1963) have yielded similar results. Readministration of the test to groups of kindergarten children on consecutive days revealed no significant difference in performance on different days. Examiner effect was also found to be negligible, as was the effect of art training in school. The old and new scales are apparently quite similar; their scores correlate between .91 and .98 in homogeneous age groups. The correlation of the Man and Woman scales is about as high as the split-half reliability of the Man scale found in comparable samples. On this basis, Harris recommends that the two scales be regarded as alternate forms and that the mean of their standard scores be used for greater reliability. The Quality scales, representing a quicker but cruder scoring method, yield interscorer reliabilities clustering in the .80's. Correlations of about the same magnitude have been found between Quality scale ratings and point scores obtained for the same drawings.

Apart from the item-analysis data gathered in the development of the scales, information regarding the construct validity of the test is provided by correlations with other intelligence tests. These correlations vary widely, but the majority are over .50. In a study with 100 fourth-grade children, correlations were found between the Draw-a-Man Test and a number of tests of known factorial composition (Ansbacher, 1952). Such correlations indicated that, within the ages covered, the Draw-a-Man Test correlates highest with tests of reasoning, spatial aptitude, and perceptual accuracy. Motor coordination plays a negligible role in the test at these ages. For kindergarten children, the Draw-a-Man Test correlated higher with numerical aptitude and lower with perceptual speed and accuracy than it did for fourth-grade children (D. B. Harris, 1963). Such findings suggest that the test may measure somewhat different functions at different ages.

The original Draw-a-Man Test has been administered widely in clinics as a supplement to the Stanford-Binet and other verbal scales. It has also been employed in a large number of studies on different cultural and ethnic groups, including several American Indian samples. Such investigations have indicated that performance on this test is more dependent on differences in cultural background than was originally assumed. In a review of studies pertaining to this test, Goodenough and Harris (1950, p. 399) expressed the opinion that "the search for a culture-free test, whether of intelligence, artistic ability, personal-social characteristics, or any other measurable trait is illusory." This view was reaffirmed by Harris in his 1963 book. More recently, Dennis (1966) analyzed comparative data obtained with this test in 40 widely different cultural groups, principally from 6-year-old children. Mean group scores appeared to be most closely related to the amount of experience with representational art within each culture. In the case of groups with little indigenous art, it was hypothesized that test performance reflects degree of acculturation to Western civilization.

Cultural differences in experiential background were again revealed in a well-designed comparative investigation of Mexican and American children with the Goodenough-Harris test (Laosa, Swartz, & Diaz-Guerrero, 1974). In studies of this test in Nigeria (Bakare, 1972) and Turkey (Uçman, 1972), mean scores increased consistently and significantly with the children's socioeconomic level. It should be added that these findings with the Goodenough-Harris test are typical of results obtained with all tests initially designed to be "culture-free" or "culture-fair."

APPROACHES TO CROSS-CULTURAL TESTING. Theoretically we can identify three approaches to the development of tests for persons reared in different cultures or subcultures, although in practice some features from all three may be combined. The first approach involves the choice of items common to many cultures and the validation of the resulting test against local criteria in many different cultures. This is the basic approach of the culture-fair tests, although their repeated validation in different cultures has often been either neglected altogether or inadequately executed. Without such a step, however, we cannot be sure that the test is relatively free from culturally restricted elements. Moreover, it is unlikely that any single test could be designed that would fully meet these requirements across a wide range of cultures.

On the other hand, cross-cultural assessment techniques are needed for basic research on some very fundamental questions. One of these questions pertains to the generality of psychological principles and constructs derived within a single culture (Anastasi, 1958a, Ch. 18). Another question concerns the role of environmental conditions in the development of individual differences in behavior—a problem that can be more effectively studied within the wide range of environmental variation provided by highly dissimilar cultures. Research of this sort calls for instruments that can be administered under at least moderately comparable conditions in different cultures. Safeguards against incorrect interpretations of results obtained with such instruments should be sought in appropriate experimental designs and in the investigators' thorough familiarity with the cultures or subcultures under investigation.

A second major approach is to make up a test within one culture and administer it to individuals with different cultural backgrounds. Such a procedure would be followed when the object of testing is prediction of a local criterion within a particular culture. In such a case, if the specific cultural loading of the test is reduced, the test validity may also drop, since the criterion itself is culturally loaded. On the other hand, we should avoid the mistake of regarding any test developed within a single cultural framework as a universal yardstick for measuring "intelligence." Nor should we assume that a low score on such a test has the same causal explanation when obtained by a member of another culture as when obtained by a member of the test culture. What *can* be ascertained by such an approach is the cultural distance between groups, as well as the individual's degree of acculturation and his readiness for educational and vocational activities that are culture-specific.

As a third approach, different tests may be developed within each culture and validated against local criteria only. This approach is exemplified by the many revisions of the original Binet scales for use in different European, Asian, and African cultures, as well as by the development of tests for industrial and military personnel within particular cultures. A current example is provided by the test-development program conducted in several developing nations of Africa, Asia, and Latin America by the

American Institutes for Research, under the sponsorship of the United States Agency for International Development (Schwarz, 1964a, 1964b; Schwarz & Krug, 1972). Another example is the long-term testing program of the National Institute of Personnel Research in Johannesburg (Blake, 1972). In such instances, the tests are validated against the specific educational and vocational criteria they are designed to predict, and performance is evaluated in terms of local norms. Each test is applied only within the culture in which it was developed and no cross-cultural comparisons are attempted. If the criteria to be predicted are technological, however, "Western-type intelligence" is likely to be needed, and the tests will reflect the direction in which the particular culture is evolving rather than its prevalent cultural characteristics at the time (see also Vernon, 1969, Ch. 14).

Attention should also be called to the publication, in the late 1960s and early 1970s, of several handbooks concerned with cross-cultural testing and research, and with the use of tests in developing countries (Biesheuvel, 1969; Brislin, Lonner, & Thorndike, 1974; Schwarz & Krug, 1972). All provide information on recommended tests, adaptations of standardized tests, and procedural guidelines for the development and application of tests. Further indication of the widespread interest in cross-cultural testing can be found in the report of an international conference on Mental Tests and Cultural Adaptation held in 1971 in Istanbul, Turkey (Cronbach & Drenth, 1972). The papers presented at this conference reflect the wide diversity of interests and backgrounds of the participants. The topics range from methodological problems and evaluations of specific instruments to theoretical discussions and reports of empirical studies.

The principal focus in both the handbooks and the conference report is on major cultural differences, as found among nations and among peoples at very different stages in their cultural evolution. In addition, a vast amount of literature has accumulated in the decades of the 1960s and 1970s on the psychological testing of minorities in the United States, chiefly for educational and vocational purposes. In the present book, this material is treated wherever it can be most clearly presented. Thus in Chapter 3, the focus was on social and ethical concerns and responsibilities in the use of tests with cultural minorities. The technical psychometric problems of test bias and item-group interactions were considered in Chapters 7 and 8. In the present chapter, attention was centered on instruments developed for cross-cultural ability testing. Problems in the interpretation of the results of cross-cultural testing will be considered in Chapter 12.

A final point should be reiterated about the instruments discussed in this section. Although initially developed for cross-cultural testing, several of these instruments have found a major application in the armamentarium of the clinical psychologist, both to supplement information obtained with such instruments as the Stanford-Binet and the Wechsler scales and in the testing of persons with various handicaps. This is especially true of the Goodenough-Harris Drawing Test, the Progressive Matrices, and the Arthur Adaptation of the Leiter scale.

CHAPTER 11

Group Testing

WHILE individual tests such as the Stanford-Binet and the Wechsler scales find their principal application in the clinic, group tests are used primarily in the educational system, civil service, industry, and the military services. It will be recalled that mass testing began during World War I with the development of the Army Alpha and the Army Beta for use in the United States Army. The former was a verbal test designed for general screening and placement purposes. The latter was a nonlanguage test for use with men who could not properly be tested with the Alpha owing to foreign-language background or illiteracy. The pattern established by these tests was closely followed in the subsequent development of a large number of group tests for civilian application.

Revisions of the civilian forms of both original army tests are still in use as Alpha Examination, Modified Form 9 (commonly known as Alpha 9) and as Revised Beta Examination. In the armed services, the Armed Forces Qualification Test (AFQT) is now administered as a preliminary screening instrument, followed by classification batteries developed within each service for assignment to occupational specialties. The AFQT provides a single score based on an equal number of vocabulary, arithmetic, spatial relations, and mechanical ability items.

In this chapter, major types of group tests in current use will be surveyed. First we shall consider the principal differences between group and individual tests. Then we shall discuss the characteristics of multilevel batteries designed to cover a wide age or grade range, with typical illustrations from different levels. Finally, group tests designed for use at the college level and beyond will be examined.

GROUP TESTS VERSUS INDIVIDUAL TESTS

ADVANTAGES OF GROUP TESTING. Group tests are designed primarily as instruments for mass testing. In comparison with individual tests, they have both advantages and disadvantages. On the positive side, group tests can be administered simultaneously to as many persons as can be fitted comfortably into the available space and reached through a microphone. Large-scale testing programs were made possible by the development of group testing techniques. By utilizing only printed items and simple responses that can be recorded on a test booklet or answer sheet, the need for a one-to-one relationship between examiner and examinee was eliminated.

A second way in which group tests facilitated mass testing was by greatly simplifying the examiner's role. In contrast to the extensive training and experience required to administer the Stanford-Binet, for example, most group tests require only the ability to read simple instructions to the examinees and to keep accurate time. Some preliminary training sessions are desirable, of course, since inexperienced examiners are likely to deviate inadvertently from the standardized procedure in ways that may affect test results. Because the examiner's role is minimized, however, group testing can provide more uniform conditions than does individual testing. The use of tapes, records, and film in test administration offers further opportunities for standardizing procedure and eliminating examiner variance in large-scale testing.

Scoring is typically more objective in group testing and can be done by a clerk. Most group tests can also be scored by computers through several available test-scoring services. Moreover, whether hand-scored or machine-scored, group tests usually provide separate answer sheets and reusable test booklets. Since in these tests all responses are written on the answer sheet, the test booklets can be used indefinitely until they wear out, thereby effecting considerable economy. Answer sheets also take up less room than test booklets and hence can be more conveniently filed for large numbers of examinees.

From another angle, group tests characteristically provide better established norms than do individual tests. Because of the relative ease and rapidity of gathering data with group tests, it is customary to test very large, representative samples in the standardization process. In the most recently standardized group tests, it is not unusual for the normative samples to number between 100,000 and 200,000, in contrast to the 2,000 to 4,000 cases laboriously accumulated in standardizing the most carefully developed individual intelligence scales.

Group tests necessarily differ from individual tests in form and arrangement of items. Although open-ended questions calling for free responses could be used—and were used in the early group tests—today the typical group test employs *multiple-choice items*. This change was obviously required for uniformity and objectivity of scoring, whether by hand or machine. With regard to arrangement of items, whereas the Binet type of scale groups items into age levels, group tests characteristically group items of similar content into *separately timed subtests*. Within each subtest, items are usually arranged in increasing order of difficulty. This arrangement ensures that each examinee has an opportunity to try each type of item (such as vocabulary, arithmetic, spatial, etc.) and to complete the easier items of each type before trying the more difficult ones on which he might otherwise waste a good deal of time.

A practical difficulty encountered with separate subtests, however, is that the less experienced or less careful examiners may make timing errors. Such errors are more likely to occur and are relatively more serious with several short time limits than with a single long time limit for the whole test. To reconcile the use of a single time limit with an arrangement permitting all examinees to try all types of items at successively increasing difficulty levels, some tests utilize the *spiral-omnibus format*. One of the earliest tests to introduce this format was the Otis Self-Administering Tests of Mental Ability which, as its name implies, endeavored to reduce the examiner's role to a minimum. The same arrangement is followed in the Otis-Lennon Mental Ability Test, from the fourth-grade level up. In a spiral-omnibus test, the easiest items of each type are presented first, followed by the next easiest of each type, and so on in a rising spiral of difficulty level, as illustrated below:

1. The opposite of hate is: Answer
    1. enemy, 2. fear, 3. love, 4. friend, 5. joy .... ( )
2. If 3 pencils cost 25 cents, how many pencils can be bought for 75 cents? .... ( )
3. A bird does not always have:
    1. wings, 2. eyes, 3. feet, 4. a nest, 5. a bill .... ( )
4. The opposite of honor is:
    1. glory, 2. disgrace, 3. cowardice, 4. fear, 5. defeat ..... ( )

In order to avoid the necessity of repeating instructions in each item and to reduce the number of shifts in instructional set required of the examinees, some tests apply the spiral-omnibus arrangement not to single items but to blocks of 5 to 10 items. This practice is followed, for example, in the Armed Forces Qualification Test and in the Scholastic Aptitude Test of the College Entrance Examination Board.

DISADVANTAGES OF GROUP TESTING. Although group tests have several desirable features and serve a well-nigh indispensable function in present-day testing, their limitations should also be noted. In group testing, the examiner has much less opportunity to establish rapport, obtain cooperation, and maintain the interest of examinees. Any temporary condition of the examinee, such as illness, fatigue, worry, or anxiety, that may interfere with test performance is less readily detected in group than in individual testing. In general, persons unaccustomed to testing may be somewhat more handicapped on group than on individual tests. There is also some evidence suggesting that emotionally disturbed children may perform better on individual than on group tests (Bower, 1969; Willis, 1970).

From another angle, group tests have been attacked because of the restrictions imposed on the examinee's responses. Criticisms have been directed particularly against multiple-choice items and against such standard item types as analogies, similarities, and classification (Hoffman, 1962; LaFave, 1966). Some of the arguments are ingenious and provocative. One contention is that such items may penalize a brilliant and original thinker who sees unusual implications in the answers. It should be noted parenthetically that if this happens, it must be a rare occurrence, in view of the item analysis and validity data. Some critics have focused on the importance of analyzing errors and inquiring into the reasons why an individual chooses a particular answer, as in the typical Piagetian approach (Sigel, 1963). It is undoubtedly true that group tests provide little or no opportunity for direct observations of the examinee's behavior or for identifying the causes of atypical performance. For this and other reasons, when important decisions about individuals are to be made, it is desirable to supplement group tests either with individual examination of doubtful cases or with additional information from other sources.

Still another limitation of traditional group testing is its lack of flexibility, insofar as every examinee is ordinarily tested on all items. Available testing time could be more effectively utilized if each examinee concentrated on items appropriate to his ability level. Moreover, such a procedure would avoid boredom from working on too easy items, at one extreme, and mounting frustration and anxiety from attempting items beyond the individual's present ability level, at the other. It will be recalled that in some individual tests, such as the Stanford-Binet and the Peabody Picture Vocabulary Test, the selection of items to be presented by the examiner depends upon the examinee's prior responses. Thus in these tests the examinee is given only items within a difficulty range appropriate to his ability level.

COMPUTER UTILIZATION IN GROUP TESTING. In the effort to combine some of the advantages of individual and group testing, several innovative techniques are being explored. Major interest thus far has centered on ways of adjusting item coverage to the response characteristics of individual examinees. In the rapidly growing literature on the topic, this approach has been variously designated as adaptive, sequential, branched, tailored, individualized, programed, dynamic, or response-contingent testing (Baker, 1971; Glaser & Nitko, 1971; Weiss & Betz, 1973). Although it is possible to design paper-and-pencil group tests that incorporate such adaptive procedures (Cleary, Linn, & Rock, 1968; Lord, 1971a), these techniques lend themselves best to computerized test administration.

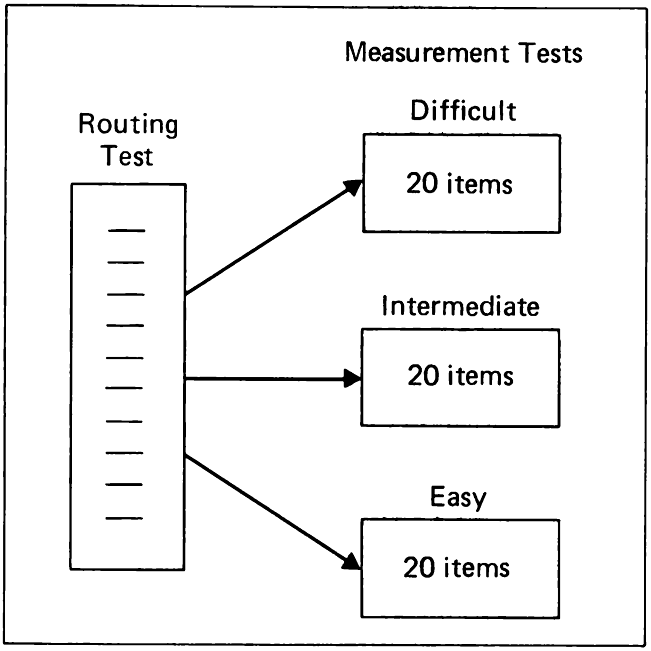
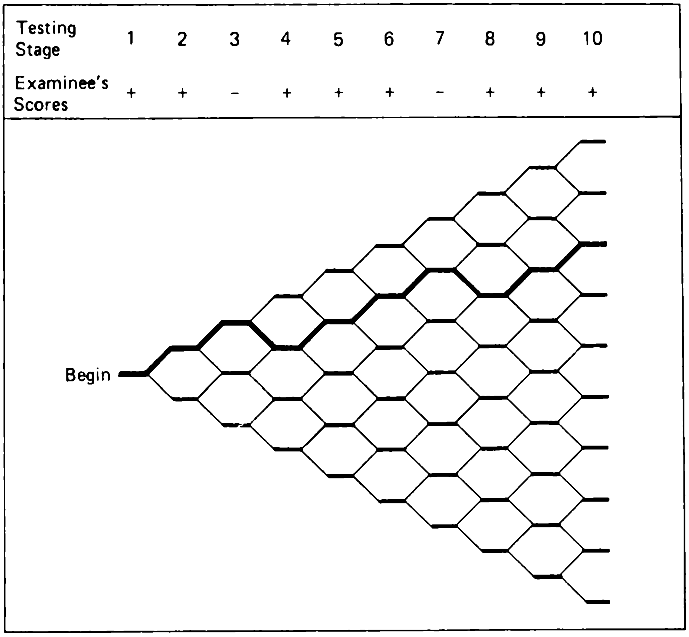


FIG. 44. Two-Stage Adaptive Testing with Three Measurement Levels. Each examinee takes routing test and one measurement test.

Adaptive testing can follow a wide variety of procedural models (DeWitt & Weiss, 1974; Larkin & Weiss, 1975; Weiss, 1974; Weiss & Betz, 1973). A simple example involving two-stage testing is illustrated in Figure 44. In this hypothetical test, all examinees take a 10-item routing test, whose items cover a wide difficulty range. Depending on his performance on the routing test, each examinee is directed to one of the three 20-item measurement tests at different levels of difficulty. Thus each person takes only 30 items, although the entire test comprises 70 items. A different arrangement is illustrated in the pyramidal test shown in Figure 45. In this case, all examinees begin with an item of intermediate difficulty. If an individual's response to this item is correct, he is routed upward to the next more difficult item; if his response is wrong, he moves downward to the next easier item. This procedure is repeated after each item response, until the individual has given 10 responses. The illustra-



Testing Stage	1	2	3	4	5	6	7	8	9	10
Examinee's Scores	+	+	−	+	+	+	−	+	+	+

FIG. 45. Pyramidal Testing Model. Heavy line shows route of examinee whose item scores are listed across top.

tion shows a 10-stage test, in which each examinee is given 10 items out of the total pool of 55 items in the test. The heavy line shows the route followed by one examinee whose responses are listed as + or − along the top.

Several variants of both of these adaptive testing models have been tried, with both paper-and-pencil and computer forms. More complex models, which do not use preestablished, fixed patterns of item sequences, are feasible only when a computer is available.[1] In general, research by various approaches indicates that adaptive tests can achieve the same reliability and validity as conventional tests in much less testing time (Larkin & Weiss, 1974; Weiss & Betz, 1973). Adaptive testing also provides greater precision of measurement for individuals at the upper and lower extremes of the ability range covered by the test (Lord, 1970, 1971b, 1971c; Weiss & Betz, 1973). In addition, adaptive testing is especially appropriate for use in the individualized instructional programs cited in Chapter 4, in which students proceed at their own pace and may therefore require test items at widely different difficulty levels. Computerized testing makes it possible to stop testing as soon as the

[1] For examples of diverse variants and models, see Weiss and Betz (1973), DeWitt and Weiss (1974), and Weiss (1974).

individual's responses provide enough information for a decision about his content mastery.

Besides providing the opportunity to adapt testing to the abilities and needs of individual examinees, computers can help to circumvent other limitations of traditional group tests(Baker, 1971; Glaser & Nitko, 1971; B. F. Green, 1970). One potential contribution is the analysis of wrong responses, in order to identify types of error in individual cases. Another is the use of response modes that permit the examinee to try alternative responses in turn, with immediate feedback, until the correct response is chosen. Still another possibility is the development of special response procedures and item types to investigate the examinee's problem-solving techniques. For instance, following the initial presentation of the problem, the examinee may have to ask the computer for further information needed to proceed at each step in the solution; or he may be required to respond by indicating the steps he follows in arriving at the solution. Considerable research is also in progress with several relatively unrestricted response modes, such as underlining the appropriate words in sentences or constructing single-word responses.

MULTILEVEL BATTERIES

OVERVIEW. Unlike individual scales of the Binet type and computerized or other adaptive tests, traditional group tests present the same items to all examinees, regardless of their individual responses. For this reason, any given group test must cover a relatively restricted range of difficulty, suitable for the particular age, grade, or ability level for which it is designed. To provide comparable measures of intellectual development over a broader range, series of overlapping multilevel batteries have been constructed. Thus, any one individual is examined only with the level appropriate to him, but other levels can be used for retesting him in subsequent years or for comparative evaluations of different age groups. The fact that successive batteries overlap provides adequate floor and ceiling for individuals at the extremes of their own age or grade distributions. It should be recognized, of course, that the match between item difficulty and examinee ability provided by multilevel batteries is at best approximate. Unlike the individualized procedures described in the preceding section, moreover, the match is based on prior knowledge about the examinee, such as his age or school grade, rather than on his own test responses.

Multilevel batteries are especially suitable for use in the schools, where comparability over several years is highly desirable. For this reason the levels are typically described in terms of grades. Most multilevel batteries provide a reasonable degree of continuity with regard to content or intellectual functions covered. Scores are expressed in terms of the same scale of units throughout. The normative samples employed at different levels are also more nearly equivalent than would be true of independently standardized tests.

Table 24 contains a list of representative multilevel batteries, together with the range of grades covered by each level. It will be noted that the individual ranges are quite narrow, most levels covering two to three grades. The total range that can be uniformly tested with a given multilevel battery, on the other hand, may extend from kindergarten to college.

The names of the batteries are also of interest. Such terms as "intelligence," "general ability," "mental ability," "mental maturity," "learning potential," "school-and-college ability," and "educational ability," are used to designate essentially the same type of test. In the psychometrician's vocabulary, these terms are virtually synonymous and interchangeable. It is noteworthy that, in the most recently developed or revised batteries, the term "intelligence" has been replaced by more specific designations. This change reflects the growing recognition that the term "intelligence" has acquired too many excess meanings, which may lead to misinterpretations of test scores. One of the newest batteries listed in Table 24, entitled "Analysis of Learning Potential", is explicitly directed

TABLE 24
Representative Multilevel Batteries

Battery	Grade Levels
Analysis of Learning Potential (ALP)	1, 2–3, 4–6, 7–9, 10–12
California Test of Mental Maturity, 1963 Revision (CTMM)	K–1, 1.5–3, 4–6, 7–9, 9–12, 12–16
Cognitive Abilities Test (CAT)	K–1, 2–3, 3, 4, 5, 6, 7, 8–9, 10–11, 12
Henmon-Nelson Tests of Mental Ability, 1973 Revision	K–2, 3–6, 6–9, 9–12
Kuhlmann-Anderson Tests: A Measure of Academic Potential	K, 1, 2, 3–4, 4–5, 5–7, 7–9, 9–13
Otis-Lennon Mental Ability Test	K, 1–1.5, 1.6–3.9, 4–6, 7–9, 10–12
School and College Ability Tests—SCAT Series II	4–6, 7–9, 10–12, 12–14
SRA Short Test of Educational Ability	K–1, 2–3, 4–6, 7–8, 9–12

toward the prediction of academic performance. Utilizing several ingenious new item types, the development of this battery began with a detailed analysis of the abilities required by school learning tasks. It thus reflects a growing emphasis on the measurement of prerequisite intellectual skills for school work and other activities of daily life. Its primary function is to assess the individual's readiness for school learning at each stage in the educational process.

Most of the batteries listed in Table 24 provide deviation IQ's or similar standard scores. Some batteries furnish several types of norms, including percentiles, stanines, or grade equivalents, as well as deviation IQ's. In addition to a total, global score, most batteries also yield separate verbal and quantitative, or linguistic and nonlinguistic scores. This breakdown is in line with the finding that an individual's performance in verbal and in other types of subtests may be quite dissimilar, especially at the upper levels. Attempts have also been made (e.g., CTMM) to report norms for the interpretation of scores on separate subtests, or combinations of subtests, representing narrower breakdowns. This practice is not to be recommended, however, because these part scores are usually too unreliable and too highly intercorrelated to permit meaningful interpretation of intraindividual differences. In general, the types of tests discussed in this chapter are suitable for assessing general intellectual development rather than relative standing in different aptitudes.

To illustrate the scope of current multilevel batteries, a different level will be examined from each of three batteries. The levels chosen are typical of those suitable for primary, elementary, and high school grades, respectively.

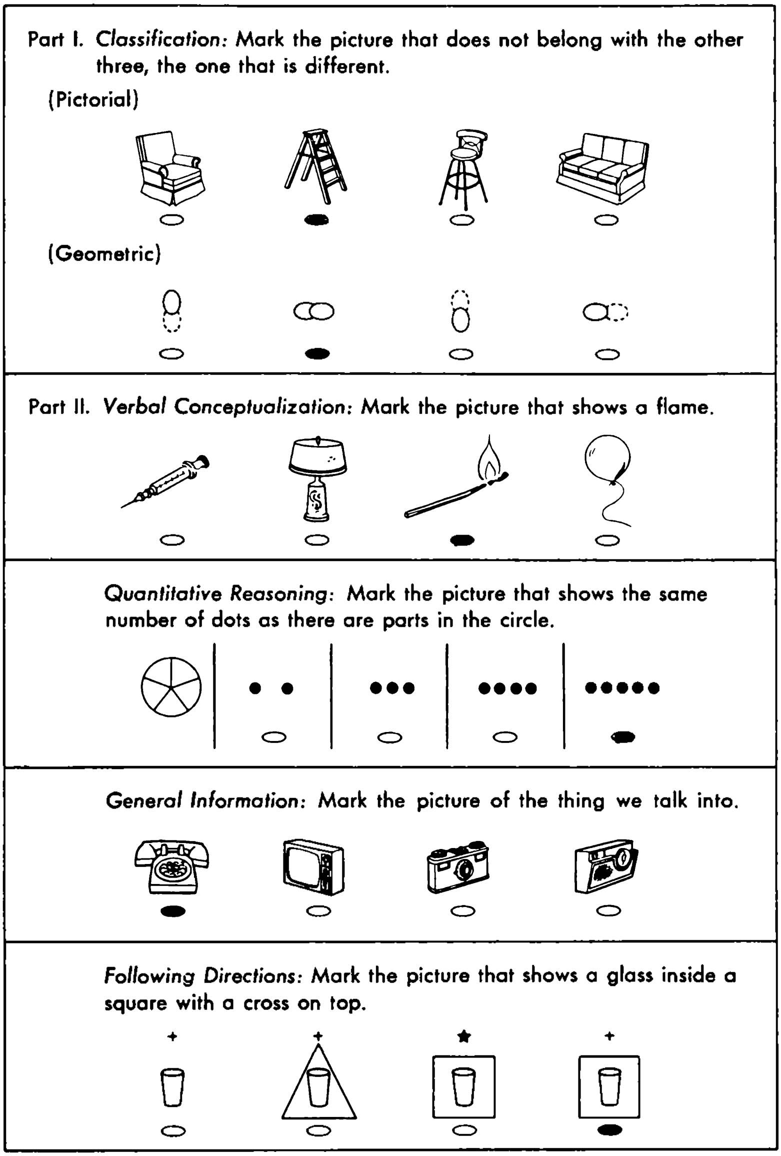
PRIMARY LEVEL. The youngest age at which it has proved feasible to employ group tests is the kindergarten and first-grade level. At the preschool ages, individual testing is required in order to establish and maintain rapport, as well as to administer the oral and performance type of items suitable for such children. By the age of 5 or 6, however, it is possible to administer printed tests to small groups of no more than 10 or 15 children. In such testing, the examiner must still give considerable individual attention to the children to make sure that directions are followed, see that pages are turned properly in the test booklets, and supervise other procedural details. With one or two assistant examiners, somewhat larger groups may be tested if necessary.

Group tests for the primary level generally cover kindergarten and the first two or three grades of elementary school. In such tests, each child is provided with a booklet on which are printed the pictures and diagrams constituting the test items. All instructions are given orally and are usually accompanied by demonstrations. Fore-exercises are frequently included in which examinees try one or two sample items and the examiner or proctor checks the responses to make certain that the instructions were properly understood. The child marks his responses on the test booklet with a crayon or soft pencil. Most of the tests require only marking the correct picture out of a set. A few call for simple motor coordination, as in drawing lines that join two dots. Obviously tests for the primary level can require no reading or writing on the part of the examinee.

Reference to Table 24 shows that several of the batteries listed have tests suitable for the primary level. To illustrate the nature of tests at this level, we shall examine the Primary Level of the Otis-Lennon Mental Ability Test. The current edition of this test, published in 1967–1969, is available in two equivalent forms (J and K) at all levels. The Primary Level actually consists of two levels, Primary I for kindergarten and Primary II for the first half of grade one. These two levels are identical in content and differ only in the way the child indicates his responses. In Primary I, responses are recorded by encircling the correct alternative; for this reason, the test must be hand scored. In Primary II, a machine-scorable booklet is used, on which responses are indicated by filling in a small oval under the correct alternative (see Fig. 46).

For each item in the entire test, the examiner gives the instructions orally. By so doing, the examiner also controls the amount of time available to complete each item (about 15 seconds). The whole test requires approximately 25 to 30 minutes and is administered in two parts with a short rest period in between. Part I consists of 23 classification items; Part II contains a total of 32 items, designed to measure verbal conceptualization, quantitative reasoning, general information, and ability to follow directions. Typical items in each of these categories are shown in Figure 46.

Norms for all levels of the Otis-Lennon battery were obtained on a carefully chosen representative sample of over 200,000 pupils in 100 school systems drawn from all 50 states. Scores can be expressed as deviation IQ's with an *SD* of 16. Percentile ranks and stanines can also be found with reference to both age and grade norms. Well-constructed tests for the primary level have generally been found to have satisfactory reliability and criterion-related validity. The Otis-Lennon Primary II yielded an alternate-form reliability of .87 in a sample of 1,047 first-grade children, over an interval of two weeks. Split-half reliability in the total sample of 14,044 first-grade children was .90. A follow-up of 144 first-grade children yielded a correlation of .80 with a higher level of the battery administered a year later. Both concurrent and predictive validity coefficients against achievement test scores and end-of-year grades cluster in the .50's.



Part I. *Classification:* Mark the picture that does not belong with the other three, the one that is different.

Part II. *Verbal Conceptualization:* Mark the picture that shows a flame.

Quantitative Reasoning: Mark the picture that shows the same number of dots as there are parts in the circle.

General Information: Mark the picture of the thing we talk into.

Following Directions: Mark the picture that shows a glass inside a square with a cross on top.

FIG. 46. Items Illustrative of the Otis-Lennon Mental Ability Test, Primary I and Primary II Levels.

(Copyright, 1967, Harcourt Brace Jovanovich, Inc.)

ELEMENTARY SCHOOL LEVEL. Group intelligence tests designed for use from the fourth grade of elementary school through high school have much in common in both content and general design. Since functional literacy is presupposed at these levels, the tests are predominantly verbal in content; most also include arithmetic problems or other numerical tests. In addition, a few batteries provide nonreading tests designed to assess the same abstract reasoning abilities in children with foreign language background, reading disabilities, or other educational handicaps.

As an example of tests for the elementary school grades, we shall consider some of the intermediate levels of the Cognitive Abilities Test. The entire series includes two primary levels (for kindergarten through the third grade) and the Multi-Level Edition covering grades 3 through 12. The Multi-Level Edition comprises eight levels (A-H) printed in a single booklet. Examinees taking different levels start and stop with different sets of items, or modules. The test is designed so that most examinees will be tested with items that are at intermediate difficulty for them, where discrimination will be most effective.

All eight levels of the Multi-Level Edition contain the same 10 subtests, grouped into three batteries as follows:

Verbal Battery—Vocabulary, Sentence Completion, Verbal Classification, Verbal Analogies.

Quantitative Battery—Quantitative Relations, Number Series, Equation Building.

Nonverbal Battery—Figure Classification, Figure Analogies, Figure Synthesis. These subtests use no words or numbers, but only geometric or figural elements; the items bear relatively little relation to formal school instruction.

Each subtest is preceded by practice exercises, the same set being used for all levels. In Figures 47, 48, and 49 will be found a typical item from each of the 10 subtests, with highly condensed instructions. In difficulty level, these items correspond to items given in grades 4 to 6. The authors recommend that all three batteries be given to each child, in three testing sessions. For most children, the Nonverbal Battery does not predict school achievement as well as do the Verbal and Quantitative batteries. However, a comparison of performance on the three batteries may provide useful information regarding special abilities or disabilities.

The standardization sample, including approximately 20,000 cases in each of the 10 grade groups, was carefully chosen to represent the school population of the country. Raw scores on each battery are translated

1. *Vocabulary:* find the one word that means the same or nearly the same as the word in dark type at the beginning of the line.

   impolite A unhappy B angry C faithless D rude E talkative

2. *Sentence Completion:* pick the one word that best fits the empty space in the sentence.

   Mark was very fond of his science teacher, but he did not ________ his mathematics teacher.

   A obey B discuss C regard D desire E like

3. *Verbal Classification:* think in what way the words in dark type go together. Then find the word in the line below that goes with them.

   dove **hawk** **sparrow**

   A moth B bat C gull D bee E squirrel

4. *Verbal Analogies:* figure out how the first two words are related to each other. Then from the five words on the line below find the word that is related to the third word in the same way.

   pea → bean : peach →

   A pit B tree C eat D skin E apple

FIG. 47. Typical Items from Verbal Battery of Cognitive Abilities Test. Answers: 1—D, 2—E, 3—C, 4—E.

(Reproduced by courtesy of Robert L. Thorndike and Elizabeth Hagen.)

into a uniform scale across all levels to permit continuity of measurement and comparability of scores in different grades. For normative interpretations, the scores on each battery can be expressed as normalized standard scores within each age, with a mean of 100 and an *SD* of 16. Percentiles and stanines can also be found, within ages and within grades. The manual advises against combining scores from the three batteries into a single index.

Kuder-Richardson reliabilities of the three battery scores, computed within grades, are mostly in the .90's. The manual also reports standard errors of measurement for different grades and score levels, as well as the minimum interbattery score differences that can be considered to have statistical and practical significance. Intercorrelations of battery scores are in the high .60's and .70's; intercorrelations of subtests are also

1. *Quantitative Relations:* If the amount or quantity in Column I is more than in Column II, mark A; if it is less, mark B; if they are equal, mark C.

Column I	*Column II*
5 x 0	5

2. *Number Series:* the numbers at the left are in a certain order. Find the number at the right that should come next.

18 16 14 12 10 → A 7 B 8 C 9 D 10 E 12

3. *Equation Building:* Arrange the numbers and signs below to make true equations and then choose the number at the right that gives you a correct answer.

1 8 9 + - A 0 B 3 C 8 D 9 E 18

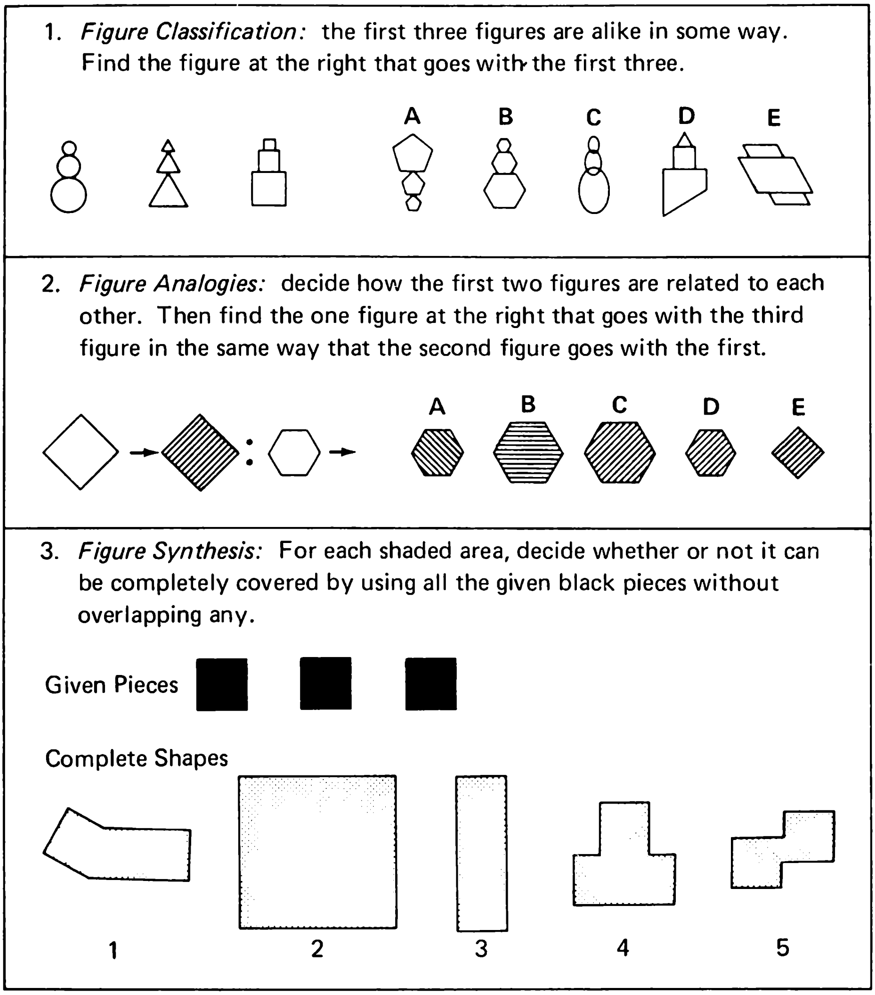
FIG. 48. Typical Items from Quantitative Battery of Cognitive Abilities Test. Answers: 1—**B**, 2—**B**, 3—**A**.

(Reproduced by courtesy of Robert L. Thorndike and Elizabeth Hagen.)

unusually high. Factor analyses likewise showed the presence of a large general factor through the three batteries, probably representing principally the ability to reason with abstract and symbolic content.

The Cognitive Abilities Test was standardized on the same normative sample as two achievement batteries, the Iowa Tests of Basic Skills (ITBS) for grades 3 to 8 and the Tests of Academic Progress (TAP) for grades 9 to 12. Concurrent validity of the Cognitive Abilities Test against the ITBS, found within single-grade groups of 500 cases from the standardization sample, ranged from the .50's to the .70's. As is generally found with academic criteria, the Verbal Battery yielded the highest correlations with achievement in all school subjects, except for arithmetic which tended to correlate slightly higher with the Quantitative Battery. Correlations with the Nonverbal Battery were uniformly lower than with the other two batteries.

Correlations with achievement tests over a three-year interval are of about the same magnitude as the concurrent correlations. Predictive validity coefficients against school grades obtained from one to three years later run somewhat lower, falling mostly in the .50's and .60's. These correlations are probably attentuated by unreliability and other extraneous variance in grading procedures.



1. *Figure Classification:* the first three figures are alike in some way. Find the figure at the right that goes with the first three.

2. *Figure Analogies:* decide how the first two figures are related to each other. Then find the one figure at the right that goes with the third figure in the same way that the second figure goes with the first.

3. *Figure Synthesis:* For each shaded area, decide whether or not it can be completely covered by using all the given black pieces without overlapping any.

FIG. 49. Typical Items from Nonverbal Battery of Cognitive Abilities Test. Answers: 1—B, 2—D, 3—3 & 4.

(Reproduced by courtesy of Robert L. Thorndike and Elizabeth Hagen and with permission of Houghton Mifflin Company.)

HIGH SCHOOL LEVEL. It should be noted that the high school levels of multilevel batteries, as well as other tests designed for high school students, are also suitable for testing general, unselected adult groups. Another source of adult tests is to be found in the tests developed for military personnel and subsequently published in civilian editions. Short screening tests for job applicants will be considered in Chapter 15.

An example of a group intelligence test for the high school level is provided by Level 2 of the School and College Ability Tests (SCAT)—Series II, designed for grades 9 to 12. At all levels of the SCAT series, tests are available in two equivalent forms, A and B. Oriented specifically toward the prediction of academic achievement, all levels yield a verbal, a quantitative, and a total score. The verbal score is based on a verbal

Part I – *Verbal Ability:* each item begins with two words that go together in a certain way. Find the two other words that go together in about the same way.

1 tool : hammer ::
A table : chair
B toy : doll
C weapon : metal
D sleigh : bell

2 braggart : humility ::
A traitor : repentance
B radical : conventionality
C precursor : foresight
D sophisticate : predisposition

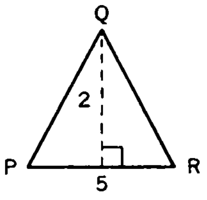
Part II – *Mathematical Ability:* each item is made up of two amounts or quantities, one in Column A and one in Column B. You have four choices: A, B, C, or D. Choose A if the quantity in Column A is greater, B if that in Column B is greater, C if the quantities are equal, or D if there is not enough information for you to tell about their sizes.

	Column A	Column B
3	A number between 10 and 20	A number between 10 and 30

	Column A	Column B
4	$\frac{1}{1000}$	$\frac{1}{1001}$

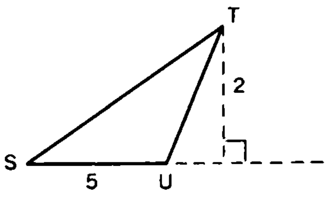
5

Column A



Area of △ PQR above

Column B



Area of △ STU above

Fig. 50. Typical Items from SCAT Series II, Level 2, for Grades 9 to 12. Answers: 1—B, 2—B, 3—D, 4—A, 5—C.

(From Student Bulletin—SCAT—Series II. Copyright © 1967 by Educational Testing Service. All rights reserved. Reprinted by permission.)

analogies test. The items in this test, however, differ from the traditional analogies items in that the respondent must choose both words in the second pair, rather than just the fourth word. The quantitative score is derived from a quantitative comparison test designed to assess the examinee's understanding of fundamental number operations. Covering both numerical and geometric content, these items require a minimum of reading and emphasize insight and resourcefulness rather than traditional computational procedures. Figure 50 shows sample verbal and quantitative items taken from a *Student Bulletin* distributed to SCAT examinees for preliminary orientation purposes. The items reproduced in Figure 50 fall approximately in the difficulty range covered by Level 2. At all levels, testing time is 40 minutes, 20 minutes for each part.

In line with current trends in testing theory, SCAT undertakes to measure developed abilities. This is simply an explicit admission of what is more or less true of all intelligence tests, namely that test scores reflect the nature and amount of schooling the individual has received rather than measuring "capacity" independently of relevant prior experiences. Accordingly, SCAT draws freely on word knowledge and arithmetic processes learned in the appropriate school grades. In this respect, SCAT does not really differ from other intelligence tests, especially those designed for the high school and college levels; it only makes overt a condition sometimes unrecognized in other tests.

Verbal, quantitative, and total scores from all SCAT levels are expressed on a common scale which permits direct comparison from one level to another. These scores can in turn be converted into percentiles or stanines for the appropriate grade. A particularly desirable feature of SCAT scores is the provision of a *percentile band* in addition to a single percentile for each obtained score. Representing a distance of approximately one standard error of measurement on either side of the corresponding percentile, the percentile band gives the 68 percent confidence interval, or the range within which are found 68 percent of the cases in a normal curve. In other words, if we conclude that an individual's true score lies within the given percentile band, the chances of our being correct are 68 out of 100 (roughly 2:1 ratio). As explained in Chapter 5, the error of measurement provides a concrete way of taking the reliability of a test into account when interpreting an individual's score.

If two percentile bands overlap, the difference between the scores can be ignored; if they do not overlap, the difference can be regarded as significant. Thus, if two students were to obtain total SCAT scores that fall in the percentile bands 55–68 and 74–84, we could conclude with fair confidence that the second actually excels the first and would continue to do so on a retest. Percentile bands likewise help in comparing a single individual's relative standing on verbal and quantitative parts of the test. If a student's verbal and quantitative scores correspond to the percentile

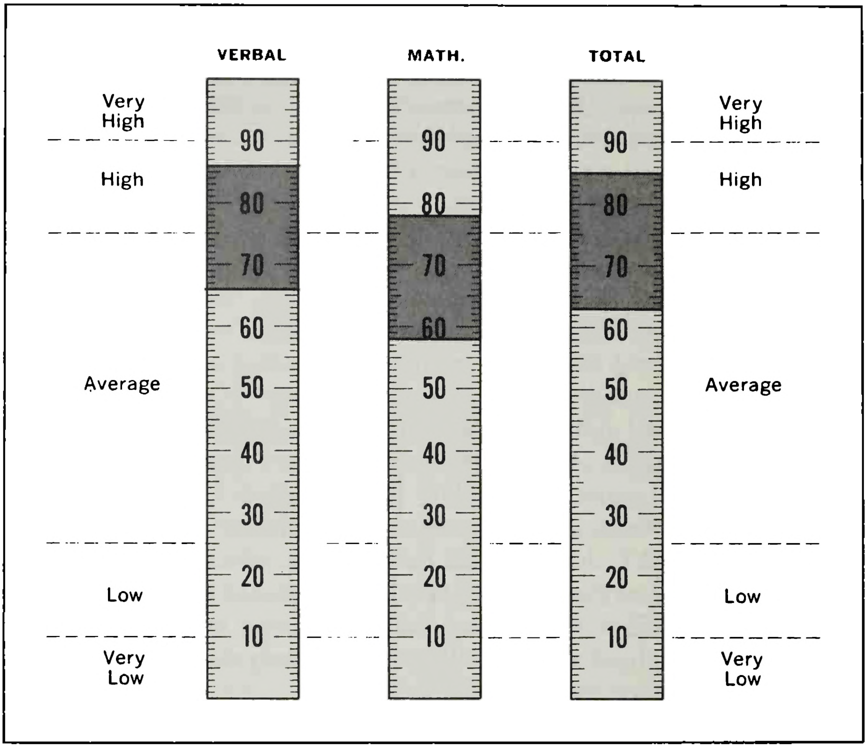


FIG. 51. SCAT—II Profile, Illustrating Percentile Bands.

(From Student Bulletin—SCAT—Series II. Copyright © 1967 by Educational Testing Service. All rights reserved. Reprinted by permission.)

bands 66–86 and 58–78, respectively, we would conclude that he is *not* significantly better in verbal than in quantitative abilities, because his percentile bands for these two scores overlap (see Fig. 51).

The SCAT standardization sample of over 100,000 cases was selected so as to constitute a representative cross-section of the national school population in grades 4 through 12 and the first two years of college. For this purpose, a three-stage sampling procedure was employed, in which the units to be sampled were school systems (public and private), schools, and classrooms, respectively. Similar successive sampling procedures were followed for the college sample. The selection of the standardization sample, as well as other test-construction procedures followed in the development of SCAT, sets an unusually high standard of technical quality.

Reliability coefficients for verbal, quantitative, and total scores were separately computed within single grade groups by the Kuder-Richardson technique. The reported reliabilities are uniformly high. Within separate grade groups, from grade 4 to 14, total score reliabilities are all .90 or above; verbal and quantitative reliabilities vary between .83 and .91. These reliabilities may be spuriously high because the tests are somewhat speeded. The percentage of students reaching the last item in different grades ranged from 65 to 96 on the verbal test and from 55 to 85 on the quantitative test. Under these circumstances, equivalent-form reliability would seem more appropriate. If the reliability coefficients are in fact spuriously high, the errors of measurement are underestimated; and hence the percentile bands should be wider.

It should be noted, however, that many of the students who did not record an answer for all items may have abandoned any attempt to solve the more difficult items even though they had enough time. In the quantitative test, moreover, a student can waste an inordinate amount of time in reaching the answer to certain items by counting or computation, when the proper recognition of numerical relations would yield the answer almost instantaneously. Under these conditions, speed of performance would be highly correlated with the quantitative reasoning abilities the test is designed to measure.

In view of the stated purpose for which SCAT was developed, its predictive validity against academic achievement is of prime relevance. Follow-up data in grades 5, 8, 11, and 12 were obtained from schools participating in the standardization sample. Validity coefficients were computed in each school and then averaged for each grade across schools, the number of available schools per grade ranging from 3 to 26. For the four grades investigated, the average correlation between SCAT Total and grade-point average ranged from .59 to .68; for SCAT Verbal and English grades, the range was .41 to .69; and for SCAT Quantitative and mathematics grades, the range was .43 to .65. Because individual correlations varied widely from school to school, however, the manual recommends local validation.

Correlations with achievement tests (Sequential Tests of Educational Progress) generally fell between .60 and .80. Quantitative scores tend to correlate higher than Verbal scores with mathematics achievement; and Verbal scores tend to correlate higher than Quantitative scores with achievement in all other subjects. Total SCAT scores, however, generally yield validity coefficients as high as those of either of the two part scores or higher. Thus the effectiveness of Verbal and Quantitative scores as differential predictors of academic achievement remains uncertain. It is noteworthy in this connection that the Verbal and Quantitative scores are themselves highly correlated. These correlations are in the .70's, except at the lowest and highest grades, where they drop to the .60's. Such close similarity may result from the item types employed in the two tests, which involve largely the ability to detect and utilize relations in symbolic and abstract content. Like other tests discussed in this chapter, SCAT was designed principally as a measure of general intellectual development and only secondarily as an indicator of intraindividual aptitude differences.

TESTS FOR THE COLLEGE LEVEL AND BEYOND

COLLEGE ADMISSION. A number of tests have been developed for use in the admission, placement, and counseling of college students. An outstanding example is the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board. Several new forms of this test are prepared each year, a different form being employed in each administration. Separate scores are reported for the Verbal and Mathematics sections of the test. Figures 52 and 53 contain brief descriptions of the verbal and mathematics item types, with illustrations. The items reproduced are among those given in the orientation booklet distributed to prospective examinees (College Entrance Examination Board, 1974b). Changes introduced in 1974 on an experimental basis include the addition of a Test of Standard Written English and the separate reporting of a vocabulary score (based on the antonyms and analogies items) and a reading com-

Antonyms: choose the lettered word or phrase which is most nearly *opposite* in meaning to the word in capital letters.

1. SCHISM: (A) majority (B) union (C) uniformity (D) conference
(E) construction

Sentence Completions: choose the one word or set of words which, when inserted in the sentence, *best* fits in with the meaning of the sentence as a whole.

2. From the first the islanders, despite an outward ____, did what they could to ____ the ruthless occupying power.

(A) harmony . . assist (B) enmity . . embarrass (C) rebellion . . foil
(D) resistance . . destroy (E) acquiescence . . thwart

Analogies: select the lettered pair which best expresses a relationship similar to that expressed in the original pair.

3. CRUTCH: LOCOMOTION: (A) paddle : canoe (B) hero : worship
(C) horse : carriage (D) spectacles : vision (E) statement : contention

Reading Comprehension: examinee reads a passage and answers multiple-choice questions assessing his understanding of its content.

FIG. 52. Illustrative Verbal Items from CEEB Scholastic Aptitude Test. Instructions have been highly condensed. Answers: 1—B, 2—E, 3—D.

(From *About the SAT 1974–75*, pp. 4, 5. College Entrance Examination Board, New York. Reprinted by permission of Educational Testing Service, copyright owner of the test questions.)

Standard Multiple-Choice Questions: drawing upon elementary arithmetic, algebra, and geometry taught in the ninth grade or earlier, these items emphasize insightful reasoning and the application of principles.

1. If 16 X 16 X 16 = 8 X 8 X P, then P =

(A) 4 (B) 8 (C) 32 (D) 48 (E) 64

Quantitative Comparisons: mark (A) if the quantity in Column A is the greater, (B) if the quantity in Column B is the greater, (C) if the two quantities are equal, (D) if the relationship cannot be determined from the information given.

	Column A	*Column B*
2.	3 X 353 X 8	4 X 352 X 6

FIG. 53. Illustrative Mathematics Items from CEEB Scholastic Aptitude Test. Instructions have been highly condensed. Answers: 1—E, 2—A.

(From *About the SAT 1974–75*, pp. 4, 5. College Entrance Examination Board, New York. Reprinted by permission of Educational Testing Service, copyright owner of the test questions.)

prehension score (based on the sentence completion and reading comprehension items).

First incorporated into the CEEB testing program in 1926, the SAT has undergone continuing development and extensive research of high technical quality. One of the reviewers in the *Seventh Mental Measurements Yearbook* writes: "Technically, the SAT may be regarded as highly perfected—possibly reaching the pinnacle of the current state of the art of psychometrics" (DuBois, 1972). Another comments: "The system of pretesting of items, analysis, and standardization of new forms exemplifies the most sophisticated procedures in modern psychometrics" (W. L. Wallace, 1972). Several aspects of the SAT research have been cited in different chapters of this book to illustrate specific procedures. A detailed account of methodology and results can be found in the technical report edited by Angoff (1971b). A shorter comparable form, known as the Preliminary SAT, has also been in use since 1959. Generally taken at an earlier stage, this test provides a rough estimate of the high school student's aptitude for college work and has been employed for educational counseling and other special purposes. Both tests are restricted to the testing program administered by the College Entrance Examination Board on behalf of member colleges. All applicants to these colleges take the SAT. Some colleges also require one or more achievement tests in special fields, likewise administered by CEEB.

Another nationwide program, launched in 1959, is the American College Testing Program (ACT). Originally limited largely to state university systems, this program has grown rapidly and is now used by many colleges throughout the country. The ACT Test Battery includes four parts: English Usage, Mathematics Usage, Social Studies Reading, and Natural Sciences Reading. Reflecting the point of view of its founder, E. F. Lindquist, the examination provides a set of work samples of college work. It overlaps traditional aptitude and achievement tests, focusing on the basic intellectual skills required for satisfactory performance in college.

Technically, the ACT does not come up to the standards set by the SAT. Reliabilities are generally lower than desirable for individual decisions. The separate scores are somewhat redundant insofar as the four parts are heavily loaded with reading comprehension and highly intercorrelated. On the other hand, validity data compare favorably with those found for other instruments in similar settings. Correlations between composite scores on the whole battery and college grade-point averages cluster around .50. Most of these validity data were obtained through research services made available to member colleges by the ACT program staff. The program also provides extensive normative and interpretive data and other ancillary services.

In addition to the above restricted tests, a number of tests designed for college-bound high school students and for college students are commercially available to counselors and other qualified users. An example is the College Qualification Tests. This battery offers six possible scores: Verbal, Numerical, Science Information, Social Studies Information, Total Information, and a Total score on the entire test. The information demanded in the various fields is of a fairly general and basic nature and is not dependent on technical aspects of particular courses. Reliability and normative data compare favorably with those of similar tests. Validity data are promising but understandably less extensive than for tests that have been used more widely.

It will be noted that, with the exception of the College Board's SAT (which can be supplemented with achievement tests), all these tests sample a combination of general aptitudes and knowledge about (or ability to handle) subject matter in major academic fields. When separate scores are available, their differential validity in predicting achievement in different fields is questionable. It would seem that total score provides the best predictor of performance in nearly all college courses. Among the part-scores, verbal scores are usually the best single predictors. Another important point to bear in mind is that scores on any of these tests are not intended as substitutes for high school grades in the prediction of college achievement. High school grades can predict college achievement as well as most tests or better. When test scores are combined with high school grades, however, the prediction of college performance is considerably improved.

GRADUATE SCHOOL ADMISSION. The practice of testing applicants for admission to college was subsequently extended to include graduate and professional schools.[2] Most of the tests designed for this purpose represent a combination of general intelligence and achievement tests. A well-known example is the Graduate Record Examinations (GRE). This series of tests originated in 1936 in a joint project of the Carnegie Foundation for the Advancement of Teaching and the graduate schools of four eastern universities. Now greatly expanded, the program is conducted by Educational Testing Service, under the general direction of the Graduate Record Examinations Board. Students are tested at designated centers prior to their admission to graduate school. The test results are used by the universities to aid in making admission and placement decisions and in selecting recipients for scholarships, fellowships, and special appointments.

The GRE include an Aptitude Test and an Advanced Test in the student's field of specialization. The latter is available in many fields, such as biology, English literature, French, mathematics, political science, and psychology. The Aptitude Test is essentially a scholastic aptitude test suitable for advanced undergraduates and graduate students. Like many such tests, it yields separate Verbal and Quantitative scores. The verbal items require verbal reasoning and comprehension of reading passages taken from several fields. The quantitative items require arithmetic and algebraic reasoning, as well as the interpretation of graphs, diagrams, and descriptive data.

Scores on all GRE tests are reported in terms of a single standard score scale with a mean of 500 and an *SD* of 100. These scores are directly comparable for all tests, having been anchored to the Aptitude Test scores of a fixed reference group of 2,095 seniors examined in 1952 at 11 colleges. A score of 500 on an Advanced Physics Test, for example, is the score expected from physics majors whose Aptitude Test score equals the mean Aptitude Test score of the reference group. Since graduate school applicants are a selected sample with reference to academic aptitude, the means of most groups actually taking each Advanced Test in the graduate student selection program will be considerably above 500. Moreover, there are consistent differences in the intellectual caliber of students majoring in different subjects. For normative interpretation, therefore, the current percentiles given for specific groups are more relevant and local norms are still better.

The reliability and validity of the GRE have been investigated in a number of different student samples (*Guide for the use of the GRE,*

---

[2] The testing of applicants to professional schools will be discussed in Chapter 15, in connection with occupational tests.

1973). Kuder-Richardson reliabilities of the Verbal and Quantitative scores of the Aptitude Test and of total scores on the Advanced Tests are consistently over .90. Several Advanced Tests also report scores in two or three major subdivisions of the field, such as experimental and social psychology. The reliabilities of these subscores are mostly in the .80's. The lower reliabilities, as well as the high intercorrelations among the parts, call for caution in the interpretation of subscores.

Predictive validity has been checked against such criteria as graduate school grade-point average, performance on departmental comprehensive examinations, overall faculty ratings, and attainment of the PhD (Willingham, 1974). In general, the GRE composite score, including Aptitude and Advanced tests, proved to be more valid than undergraduate grade-point average as a predictor of graduate school performance. Depending upon the criterion used, the difference in favor of the GRE varied from slight to substantial. Consistent with expectation, GRE–Q was a better predictor than GRE–V in those scientific fields in which mathematical ability is of major importance; the reverse was true of such fields as English. The GRE Advanced Test was the most generally valid single predictor among those investigated. Illustrative data from three fields can be seen in Figure 54, showing the percentage of students attaining the PhD in successive intervals of Advanced Test scores. The three coefficients given in Figure 54 are biserial correlations between GRE Advanced Test scores and attainment or nonattainment of the PhD.

The highest validities were obtained with weighted composites of undergraduate grade-point average and one or more GRE scores. These multiple correlations fell mostly between .40 and .45 for various criteria and for different fields. It should be noted that the narrow range of talent covered by graduate school applicants necessarily results in lower correlations than are obtained with the SAT among college applicants. This finding does not imply that the GRE is intrinsically less valid than the SAT; rather it means that finer discriminations are required within the more narrowly restricted graduate school population.

Another widely used test for the selection of graduate students is the Miller Analogies Test. Consisting of complex analogies items whose subject matter is drawn from many academic fields, this test has an unusually high ceiling. Although a 50-minute time limit is imposed, the test is primarily a power test. The Miller Analogies Test was first developed for use at the University of Minnesota, but later forms were made available to other graduate schools and it was subsequently published by the Psychological Corporation. Its administration, however, is restricted to licensed centers in universities or business organizations. The test is used both in the selection of graduate students and in the evaluation of personnel for high-level jobs in industry. It is available in five parallel forms, one of which is reserved for reexaminations.

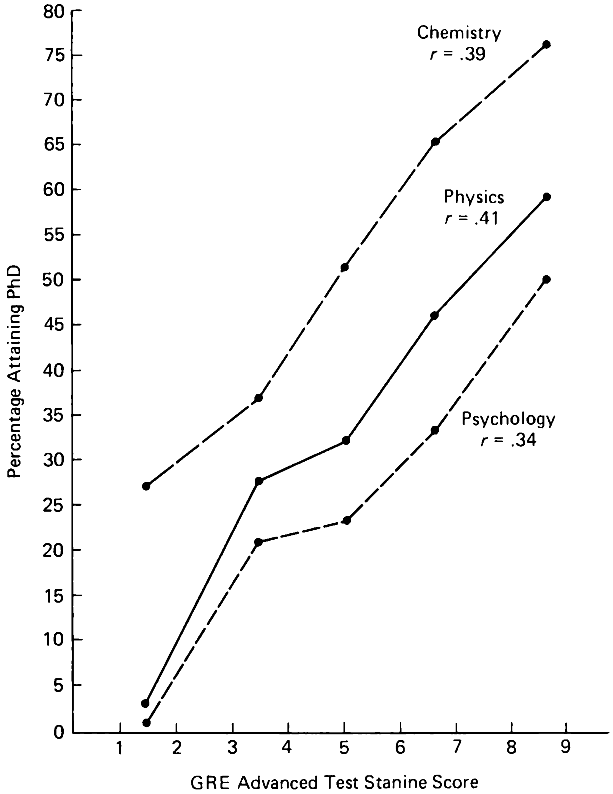


FIG. 54. Percentage of Students at Various Levels of GRE Advanced Test Scores Who Attained the PhD within 10 Years.

(From Willingham, 1974, p. 276; data from Creager, 1965. Copyright 1974 by the American Association for the Advancement of Science.)

Percentile norms on the Miller Analogies Test are given for graduate and professional school students in several fields and for groups of industrial employees and applicants. Over half of these groups contained 500 or more cases and none had less than 100. Marked variations in test performance are found among these different samples. The median of one group, for example, corresponds to the 90th percentile of another. Means and *SD*'s for additional, smaller industrial samples are also reported as further aids in normative interpretation.

Odd-even reliability coefficients of .92 to .95 were found in different samples, and alternate-form reliabilities ranged from .85 to .90. Correlations with several individual and group tests of intelligence and academic aptitudes fall almost entirely between the .50's and the .70's. Over 100 validity coefficients are reported for graduate and professional student groups and for a few industrial samples. These coefficients vary widely. Slightly over a third are between .30 and .60. About an equal number are clearly too low to be significant. The field of specialization, the nature of the criteria employed, and the size, heterogeneity, and other characteristics of the samples are obvious conditions affecting these coefficients. Means and *SD*'s of several contrasting groups in different settings provide some additional promising validity data. It is evident that the validity of this test must be evaluated in terms of the specific context in which it is to be used.

SUPERIOR ADULTS. Any test designed for college or graduate students is also likely to be suitable for examining superior adults for occupational assessment, research, or other purposes. The use of the Miller Analogies Test for the selection and evaluation of high-level industrial personnel has already been mentioned. Another test that provides sufficient ceiling for the examination of highly superior adults is the Concept Mastery Test (CMT). Originating as a by-product of Terman's extensive longitudinal study of gifted children, Form A of the Concept Mastery Test was developed for testing the intelligence of the gifted group in early maturity (Terman & Oden, 1947). For a still later follow-up, when the gifted subjects were in their mid-forties, Form T was prepared (Terman & Oden, 1959). This form, which is somewhat easier than Form A, was subsequently released for more general use.

The Concept Mastery Test consists of both analogies and synonym-antonym items. Like the Miller Analogies Test, it draws on concepts from many fields, including physical and biological sciences, mathematics, history, geography, literature, music, and others. Although predominantly verbal, the test incorporates some numerical content in the analogies items.

Percentile norms are given for approximately 1,000 cases in the Stanford gifted group, tested at a mean age of 41 years, as well as for smaller samples of graduate students, college seniors applying for Ford Foundation Fellowships in Behavioral Sciences, and engineers and scientists in a navy electronics laboratory. To provide further interpretive data, the manual (with 1973 supplement) reports means and *SD*'s of some 20 additional student and occupational samples.

Alternate-form reliabilities of the CMT range from .86 to .94. Scores show consistent rise with increasing educational level and yield correlations clustering around .60 with predominantly verbal intelligence tests. Significant correlations with grade-point averages were found in seven college samples, the correlations ranging from .26 to .59. Some suggestive findings in other contexts are also cited. For example, in two groups of managers participating in advanced management training programs, CMT scores correlated .40 and .45 with peer ratings of ability to think critically and analytically. And in a group of 200 experienced elementary and secondary school teachers, CMT correlated .54 with a scale designed to measure teachers' attitude toward gifted children. Evidently the teachers who themselves scored higher on this test had more favorable attitudes toward gifted children.

Because of its unique features, the Concept Mastery Test can undoubtedly serve a useful function for certain testing purposes. On the other hand, it is clearly not an instrument that can be used or interpreted routinely. A meaningful interpretation of CMT scores requires a careful study of all the diverse data accumulated in the manual, preferably supplemented by local norms.

CHAPTER 12

Psychological Issues in Intelligence Testing

PSYCHOLOGICAL tests should be regarded as tools. Like all tools, their effectiveness depends on the knowledge, skill, and integrity of the user. A hammer can be employed to build a crude kitchen table or a fine cabinet—or as a weapon of assault. Since psychological tests are measures of behavior, the interpretation of test results requires knowledge about human behavior. Psychological tests cannot be properly applied outside the context of psychological science. Familiarity with relevant behavioral research is needed not only by the test constructor but also by the test user.

An inevitable consequence of the expansion and growing complexity of any scientific endeavor is an increasing specialization of interests and functions among its practitioners. Such specialization is clearly apparent in the relationship of psychological testing to the mainstream of contemporary psychology (Anastasi, 1967). Specialists in psychometrics have raised techniques of test construction to truly impressive pinnacles of quality. While providing technically superior instruments, however, they have given relatively little attention to ensuring that test users had the psychological information needed for the proper use of such instruments. As a result, outdated interpretations of test performance all too often survived without reference to the results of pertinent behavioral research. This partial isolation of psychological testing from other fields of psychology—with its consequent misuses and misinterpretations of tests—accounts for some of the public discontent with psychological testing in the decades of the 1950s and 1960s. The topics chosen for discussion in this chapter illustrate ways in which the findings of psychological research can contribute to the effective use of intelligence tests and can help to correct popular misconceptions about the IQ and similar scores.

LONGITUDINAL STUDIES OF INTELLIGENCE

An important approach to the understanding of the construct, "intelligence," is through longitudinal studies of the same individuals over long periods of time. Although such investigations may be regarded as contributing to the long-term predictive validation of specific tests, they have broader implications for the nature of intelligence and the meaning of an IQ. When intelligence was believed to be largely an expression of hereditary potential, each individual's IQ was expected to remain very nearly constant throughout life. Any observed variation on retesting was attributed to weaknesses in the measuring instrument—either inadequate reliability or poor selection of functions tested. With increasing research on the nature of intelligence, however, has come the realization that intelligence itself is both complex and dynamic. In the following sections, we shall examine typical findings of longitudinal studies of intelligence and shall inquire into the conditions making for both stability and instability of the IQ.

STABILITY OF THE IQ. An extensive body of data has accumulated showing that, over the elementary, high school, and college period, intelligence test performance is quite stable (see Anastasi, 1958a, pp. 232–238; McCall, Appelbaum, & Hogarty, 1973). In a Swedish study of a relatively unselected population, for example, Husén (1951) found a correlation of .72 between the test scores of 613 third-grade school boys and the scores obtained by the same persons 10 years later on their induction into military service. In a later Swedish study, Härnqvist (1968) reports a correlation of .78 between tests administered at 13 and 18 years of age to over 4,500 young men. Even preschool tests show remarkably high correlations with later retests. In a longitudinal study of 140 children conducted at Fels Research Institute (Sontag, Baker, & Nelson, 1958), Stanford-Binet scores obtained at 3 and at 4 years of age correlated .83. The correlation with the 3-year tests decreased as the interval between retests increased, but by age 12 it was still as high as .46. Of special relevance to the Stanford-Binet is the follow-up conducted by Bradway, Thompson, and Cravens (1958) on children originally tested between the ages of 2 and 5½ as part of the 1937 Stanford-Binet standardization sample. Initial IQ's correlated .65 with 10-year retests and .59 with 25-year retests. The correlation between the 10-year retest (Mean age = 14 years) and 25-year retest (Mean age = 29 years) was .85.

As would be expected, retest correlations are higher, the shorter the interval between tests. With a constant interval between tests, moreover, retest correlations tend to be higher the older the children. The effects of age and retest interval on retest correlations exhibit considerable regularity and are themselves highly predictable (Thorndike, 1933, 1940). One explanation for the increasing stability of the IQ with age is provided by the cumulative nature of intellectual development. The individual's intellectual skills and knowledge at each age include all his earlier skills and knowledge plus an increment of new acquisitions. Even if the annual increments bear no relation to each other, a growing consistency of performance level would emerge, simply because earlier acquisitions constitute an increasing proportion of total skills and knowledge as age increases. Predictions of IQ from age 10 to 16 would thus be more accurate than from 3 to 9, because scores at 10 include over half of what is present at 16, while scores at 3 include a much smaller proportion of what is present at 9.

Anderson (1940) described this relationship between successive scores as the *overlap hypothesis.* He maintained that, "Since the growing individual does not lose what he already has, the constancy of the IQ is in large measure a matter of the part-whole or overlap relation" (p. 394). In support of this hypothesis, Anderson computed a set of correlations between initial and terminal "scores" obtained with shuffled cards and random numbers. These correlations, which depended solely on the extent of overlap between successive measures, agreed closely with empirical test-retest correlations in intelligence test scores found in three published longitudinal studies. In fact, the test scores tended to give somewhat *lower* correlations, a difference Anderson attributed to such factors as errors of measurement and change in test content with age.

Although the overlap hypothesis undoubtedly accounts for some of the increasing stability of the IQ in the developing individual, two additional conditions merit consideration. The first is the *environmental stability* characterizing the developmental years of most individuals. Children tend to remain in the same family, the same socioeconomic level, and the same cultural milieu as they grow up. They are not typically shifted at random from intellectually stimulating to intellectually handicapping environments. Hence, whatever intellectual advantages or disadvantages they had at one stage in their development tend to persist in the interval between retests.

A second condition contributing to the general stability of the IQ pertains to the role of *prerequisite learning skills* on subsequent learning. Not only does the individual retain prior learning, but much of his prior learning provides tools for subsequent learning. Hence, the more progress he has made in the acquisition of intellectual skills and knowledge at any one point in time, the better able he is to profit from subsequent learning experiences. The concept of readiness in education is an expression of this general principle. The sequential nature of learning is also implied in the previously discussed Piagetian approach to mental development, as well as in the various individualized instructional programs.

Applications of the same principle underlie Project Head Start and other compensatory educational programs for culturally disadvantaged preschool children (Bloom, Davis, & Hess, 1965; Gordon & Wilkerson, 1966; Sigel, 1973; Stanley, 1972, 1973; Whimbey, 1975). Insofar as children from disadvantaged backgrounds lack some of the essential prerequisites for effective school learning, they would only fall farther and farther behind in academic achievement as they progressed through the school grades. It should be added that learning prerequisites cover not only such intellectual skills as the acquisition of language and of quantitative concepts, but also attitudes, interests, motivation, problem-solving styles, reactions to frustration, self-concepts, and other personality characteristics.

The object of compensatory educational programs is to provide the learning prerequisites that will enable children to profit from subsequent schooling. In so doing, of course, these programs hope to disrupt the "stability" of IQ's that would otherwise have remained low. Compensatory education programs provide one example of the interaction between initial score and treatment in the prediction of subsequent score, discussed in Chapter 7. That intellectual skills can also be effectively taught at the adult level is suggested by the promising exploratory research reported by Whimbey (1975), which he describes as "cognitive therapy."

INSTABILITY OF THE IQ. Correlational studies on the stability of the IQ provide actuarial data, applicable to group predictions. For the reasons given above, IQ's tend to be quite stable in this actuarial sense. Studies of individuals, on the other hand, reveal large upward or downward shifts in IQ.[1] Sharp rises or drops in IQ may occur as a result of major environmental changes in the child's life. Drastic changes in family structure or home conditions, adoption into a foster home, severe or prolonged illness, and therapeutic or remedial programs are examples of the type of events that may alter the child's subsequent intellectual development. Even children who remain in the same environment, however, may show large increases or decreases in IQ on retesting. These changes mean, of course, that the child is developing at a faster or a slower rate than that of the normative population on which the test was standardized. In general, children in culturally disadvantaged environments tend to lose and those in superior environments to gain in IQ with age. Investigations of the specific characteristics of these environments and of the children themselves are of both theoretical and practical interest.

---

[1] See, e.g., Bayley (1955), Bayley & Schaefer (1964), Bradway (1945), Bradway & Robinson (1961), Haan (1963), Honzik, Macfarlane, & Allen (1948), Kagan & Freeman (1963), Kagan, Sontag, Baker, & Nelson (1958), McCall, Appelbaum, & Hogarty (1973), Rees & Palmer (1970), Sontag, Baker, & Nelson (1958), Wiener, Rider, & Oppel (1963).

Pinneau (1961) has prepared tables showing the median and range of individual IQ changes found in the Berkeley Growth Study for each age at test and retest from 1 month to 17 years.

Typical data on the magnitude of individual IQ changes is provided by the California Guidance Study. In an analysis of retest data from 222 cases in this study, Honzik, Macfarlane, and Allen (1948) reported individual IQ changes of as much as 50 points. Over the period from 6 to 18 years, when retest correlations are generally high, 59 percent of the children changed by 15 or more IQ points, 37 percent by 20 or more points, and 9 percent by 30 or more. Nor are most of these changes random or erratic in nature. On the contrary, children exhibit consistent upward or downward trends over several consecutive years; and these changes are related to environmental characteristics. In the California Guidance Study, detailed investigation of home conditions and parent-child relationships indicated that large upward or downward shifts in IQ were associated with the cultural milieu and emotional climate in which the child was reared. A further follow-up conducted when the participants had reached the age of 30 still found significant correlations between test scores and family milieu as assessed at the age of 21 months (Honzik, 1967). Parental concern with the child's educational achievement emerged as an important correlate of subsequent test performance, as did other variables reflecting parental concern with the child's general welfare.

In the previously mentioned follow-up of the 1937 Stanford-Binet standardization sample, Bradway (1945) selected for special study the 50 children showing the largest IQ changes from the preschool to the junior high school period. Results of home visits and interviews with parents again indicated that significant rises or drops in IQ over the 10-year period were related to various familial and home characteristics.

In a reanalysis of results obtained in five published longitudinal studies, including some of those cited directly in this chapter, Rees and Palmer (1970) found changes in IQ between 6 and 12 years to be significantly related to socioeconomic status as indicated by father's educational and occupational level. A similar relationship was observed by Härnqvist (1968) in his Swedish study. In their 10- and 25-year follow-ups of children who had taken the Stanford-Binet at preschool ages, Bradway and Robinson (1961) computed an index based on parents' education, father's occupation, and occupation of both grandfathers. Although they labeled this measure an ancestral index, rather than an index of socioeconomic status, their results are consistent with those of other investigators: the index yielded significant correlations of approximately .30

with IQ's in both follow-ups. Several longitudinal studies have demonstrated a relation between amount and direction of change in intelligence test scores and amount of formal schooling the individual himself has completed in the interval between test and retest (see Härnqvist, 1973). The score differences associated with schooling are larger than those associated with socioeconomic status of the family.

Some investigators have concentrated more specifically on the personality characteristics associated with intellectual acceleration and deceleration. At the Fels Research Institute, 140 children were included in an intensive longitudinal study extending from early infancy to adolescence and beyond (Kagan & Freeman, 1963; Kagan, Sontag, Baker, & Nelson, 1958; Sontag, Baker, & Nelson, 1958). Within this group, those children showing the largest gains and those showing the largest losses in IQ between the ages of 4½ and 6 were compared in a wide variety of personality and environmental measures; the same was done with those showing the largest IQ changes between 6 and 10. During the preschool years, emotional dependency on parents was the principal condition associated with IQ loss. During the school years, IQ gains were associated chiefly with high achievement drive, competitive striving, and curiosity about nature. Suggestive data were likewise obtained regarding the role of parental attitudes and child-rearing practices in the development of these traits.

A later analysis of the same sample, extending through age 17, focused principally on patterns of IQ change over time (McCall, Appelbaum, & Hogarty, 1973). Children exhibiting different patterns were compared with regard to child-rearing practices as assessed through periodic home visits. A typical finding is that the parents of children whose IQ's showed a rising trend during the preschool years presented "an encouraging and rewarding atmosphere, but one with some structure and enforcement of policies" (McCall et al., 1973, p. 54). A major condition associated with rising IQ's is described as accelerational attempt, or the extent to which "the parent deliberately trained the child in various mental and motor skills which were not yet essential" (p. 52).

Another approach to an understanding of IQ changes is illustrated by Haan's (1963) follow-up study of 49 men and 50 women who had participated in a long-term growth study. IQ's were obtained with a group test administered when the subjects were about 12 years old and again when they were in their middle or late 30's. Personality characteristics were investigated through a self-report inventory and a series of intensive interviews conducted at the time of the adult follow-up. The upper and lower 25 percent of the group in terms of IQ change, designated as accelerators and decelerators, were compared with special reference to their reliance on coping or defense mechanisms. These mechanisms refer to contrasting personality styles in dealing with problems and frustrations.

Coping mechanisms in general represent an objective, constructive, realistic approach; defense mechanisms are characterized by withdrawal, denial, rationalization, and distortion. The results confirmed the hypothesis that accelerators made significantly more use of coping mechanisms and decelerators of defense mechanisms.

Similar results are reported by Moriarty (1966), from a longitudinal study of 65 children tested from two to four times between infancy and the early teens. On the basis of IQ changes, the children were classified into four categories: (a) relatively constant—40 percent; (b) accelerative spurts in one or more areas of functioning—25 percent; (c) slow, delayed, or inhibited development—9 percent; (d) erratic score changes, inconsistent performance in different functions, or progressive IQ decline—26 percent. Intensive case studies of the individual children in these four categories led Moriarty to hypothesize that characteristic differences in coping mechanisms constitute a major factor in the observed course of IQ over time.

Research on the factors associated with increases and decreases in IQ throws light on the conditions determining intellectual development in general. It also suggests that prediction of subsequent intellectual status can be improved if measures of the individual's emotional and motivational characteristics and of his environment are combined with initial test scores. From still another viewpoint, the findings of this type of research point the way to the kind of intervention programs that can effectively alter the course of intellectual development in the desired directions.

INTELLIGENCE IN EARLY CHILDHOOD

The assessment of intelligence at the two extremes of the age range presents special theoretical and interpretive problems. One of these problems pertains to the functions that should be tested. What constitutes intelligence for the infant and the preschool child? What constitutes intelligence for the older adult? The second problem is not entirely independent of the first. Unlike the schoolchild, the infant and preschooler have not been exposed to the standardized series of experiences represented by the school curriculum. In developing tests for the elementary, high school, and college levels, the test constructor has a large fund of common experiential material from which he can draw test items. Prior to school entrance, on the other hand, the child's experiences are far less standardized, despite certain broad cultural uniformities in child-rearing practices. Under these conditions, both the construction of tests and the interpretation of test results are much more difficult. To some extent, the same difficulty is encountered in testing older adults, whose schooling was completed many years earlier and who have since been engaged in highly diversified activities. In this and the next section, we shall examine some of the implications of these problems for early childhood and adult testing, respectively.

PREDICTIVE VALIDITY OF INFANT AND PRESCHOOL TESTS. The conclusions that emerge from longitudinal studies is that preschool tests (especially when administered after the age of 2 years) have moderate validity in predicting subsequent intelligence test performance, but that infant tests have virtually none (Bayley, 1970; Lewis, 1973; McCall, Hogarty, & Hurlburt, 1972). Combining the results reported in eight studies, McCall and his associates computed the median correlations between tests administered during the first 30 months of life and childhood IQ obtained between 3 and 18 years (McCall et al., 1972). Their findings are reproduced in Table 25. Several trends are apparent in this table. First, tests given during the first year of life have little or no long-term predictive value. Second, infant tests show some validity in predicting IQ at preschool ages (3–4 years), but the correlations exhibit a sharp drop beyond that point, after children reach school age. Third, after the age of 18 months, validities are moderate and stable. When predictions are made from these ages, the correlations seem to be of the same order of magnitude regardless of the length of the retest interval.

TABLE 25
Median Correlations between Infant Tests and Childhood IQ

(From McCall, Hogarty, & Hurlburt, 1972. Copyright 1972 by the American Psychological Association. Reprinted by permission.)

Childhood Age in Years (Retest)	Infant Age in Months (Initial Test)			
	1–6	7–12	13–18	19–30
8–18	.01	.20	.21	.49
5–7	.01	.06	.30	.41
3–4	.23	.33	.47	.54

The lack of long-term predictive validity of infant tests needs to be evaluated further with regard to other related findings. First, a number of clinicians have argued that infant tests *do* improve the prediction of subsequent development, but only if interpreted in the light of concomitant clinical observations (Donofrio, 1965; Escalona, 1950; Knobloch & Pasamanick, 1960). Predictions might also be improved by a consideration of developmental trends through repeated testing, a procedure originally proposed by Gesell with reference to his Developmental Schedules.

In the second place, several investigators have found that infant tests have much higher predictive validity within nonnormal, clinical populations than within normal populations. Significant validity coefficients in the .60's and .70's have been reported for children with initial IQ's below 80, as well as for groups having known or suspected neurological abnormalities (Ireton, Thwing, & Gravem, 1970; Knobloch & Pasamanick, 1963, 1966, 1967; Werner, Honzik, & Smith, 1968). Infant tests appear to be most useful as aids in the diagnosis of defective development resulting from organic pathology of either hereditary or environmental origin. In the absence of organic pathology, the child's subsequent development is determined largely by the environment in which he is reared. This the test cannot be expected to predict. In fact, parental education and other characteristics of the home environment are better predictors of subsequent IQ than are infant test scores; and beyond 18 months, prediction is appreciably improved if test scores are combined with indices of familial socioeconomic status (Bayley, 1955; McCall *et al.*, 1972; Pinneau, 1961; Werner, Honzik, & Smith, 1968).

NATURE OF EARLY CHILDHOOD INTELLIGENCE. Several investigators have concluded that, while lacking predictive validity for the general population, infant intelligence tests are valid indicators of the child's cognitive abilities at the time (Bayley, 1970; Stott & Ball, 1965; Thomas, 1970). According to this view, a major reason for the negligible correlations between infant tests and subsequent performance is to be found in the changing nature and composition of intelligence with age. Intelligence in infancy is qualitatively different from intelligence at school age; it consists of a different combination of abilities.

This approach is consistent with the concept of developmental tasks proposed by several psychologists in a variety of contexts (Erikson, 1950; Havighurst, 1953; Super et al., 1957). Educationally and vocationally, as well as in other aspects of human development, the individual encounters typical behavioral demands and problems at different life stages, from infancy to senescence. Although both the problems and the appropriate reactions vary somewhat among cultures and subcultures, modal requirements can be specified within a given cultural setting. Each life stage makes characteristic demands upon the individual. Mastery of the developmental tasks of earlier stages influences the individual's handling of the behavioral demands of the next.

Within the more circumscribed area of cognitive development, Piagetian stages provide a framework for examining the changing nature of intelligence. McCall and his associates at the Fels Research Institute (McCall et al., 1972) have explored the interrelations of infant behavior in terms of such a Piagetian orientation. Through sophisticated statistical analyses involving intercorrelations of different skills within each age as well as correlations among the same and different skills across ages, these investigators looked for precursors of later development in infant behavior. Although the findings are presented as highly tentative and only suggestive, the authors describe the major component of infant intelligence at each six-month period during the first 2 years of life. These descriptions bear a rough resemblance to Piagetian developmental sequences. The main developmental trends at 6, 12, 18, and 24 months, respectively, are summarized as follows: (1) manipulation that produces contingent perceptual responses; (2) imitation of fine motor and social-vocal-verbal behavior; (3) verbal labeling and comprehension; (4) further verbal development, including fluent verbal production and grammatical maturity.

Apart from a wealth of provocative hypotheses, one conclusion that emerges clearly from the research of McCall and his co-workers is that the predominant behavior at different ages exhibits qualitative shifts and fails to support the conception of a "constant and pervasive" general intellectual ability (McCall et al., 1972, p. 746). The same conclusion was reached by Lewis (1973) on the basis of both his own research and his survey of published studies. Lewis describes infant intelligence test performance as being neither stable nor unitary. Negligible correlations may be found over intervals even as short as three months; and correlations with performance on the same or different scales at the age of two years and beyond are usually insignificant. Moreover, there is little correlation among different scales administered at the same age. These results have been obtained with both standardized instruments such as the Bayley Scales of Infant Development and with ordinal scales of the Piagetian type (Gottfried & Brody, 1975; King & Seegmiller, 1971; Lewis & McGurk, 1972). In place of the traditional model of a "developmentally constant general intelligence," Lewis proposes an interactionist view, emphasizing both the role of experience in cognitive development and the specificity of intellectual skills.

IMPLICATIONS FOR INTERVENTION PROGRAMS. The late 1960s and early 1970s witnessed some disillusionment and considerable confusion regarding the purposes, methods, and effectiveness of compensatory preschool educational programs, such as Project Head Start. Designed principally to enhance the academic readiness of children from disadvantaged backgrounds, these programs differed widely in procedures and results. Most were crash projects, initiated with inadequate planning. Only a few could demonstrate substantial improvements in the children's performance—and such improvements were often limited and short-lived (Stanley, 1972).

Against this background, the Office of Child Development of the U.S. Department of Health, Education, and Welfare sponsored a conference of a panel of experts to try to define "social competency" in early childhood (Anderson & Messick, 1974). The panel agreed that social competency includes more than the traditional concept of general intelligence. After working through a diversity of approaches and some thorny theoretical issues, the panel drew up a list of 29 components of social competency, which could serve as possible goals of early intervention programs. Including emotional, motivational, and attitudinal as well as cognitive variables, these components ranged from self-care and a differentiated self-concept to verbal and quantitative skills, creative thinking, and the enjoyment of humor, play, and fantasy. Assessment of these components requires not only a wide variety of tests but also other measurement techniques, such as ratings, records, and naturalistic observations. Few if any intervention programs could undertake to implement all the goals. But the selection of goals should be explicit and deliberate; and it should guide both intervention procedures and program evaluation.

The importance of evaluating the effectiveness of an intervention program in terms of the specific skills (cognitive or noncognitive) that the program was designed to improve is emphasized by Lewis (1973). In line with the previously cited specificity of behavioral development in early childhood, Lewis urges the measurement of specific skills rather than the use of an IQ or other broad developmental indices. Training in sensorimotor functions should not be expected to improve verbal skills; training in the development of object permanence should be assessed by a test of object permanence, and so forth. In addition, the content of the intervention program should be tailored to fit the needs of the individual child, with reference to the development attained in specific skills.

Sigel (1973) gives an incisive analysis of preschool programs against the background of current knowledge regarding both child development and educational techniques. In line with available knowledge about developmental processes, he, too, recommends the use of specific achievement tests to assess progress in the skills developed by the educational programs, instead of such global scores as IQ's. He also emphasizes the importance of process, interrelation of changes in different functions, and patterns of development, as in the characteristic Piagetian approach. And he urges the reformulation of the goals of early childhood education in more realistic terms.

PROBLEMS IN THE TESTING OF ADULT INTELLIGENCE

AGE DECREMENT. A distinctive feature introduced by the Wechsler scales for measuring adult intelligence (Ch. 9) was the use of a declining norm to compute deviation IQ's. It will be recalled that raw scores on the WAIS subtests are first transmuted into standard scores with a mean of 10 and an *SD* of 3. These scaled scores are expressed in terms of a fixed reference group consisting of the 500 persons between the ages of 20 and 34 years included in the standardization sample. The sum of the scaled scores on the 11 subtests is used in finding the deviation IQ in the appropriate age table. If we examine the sums of the scaled scores directly, however, we can compare the performance of different age groups in terms of a single, continuous scale. Figure 55 shows the means of these total scaled scores for the age levels included in the national standardization sample and for the more limited "old-age sample" of 475 persons aged 60 years and over (Doppelt & Wallace, 1955).

As can be seen in Figure 55, the scores reach a peak between the ages of 20 and 34 and then decline slowly until 60. A sharper rate of decline was found after 60 in the old-age sample. The deviation IQ is found by referring an individual's total scaled score to the norm for his own age

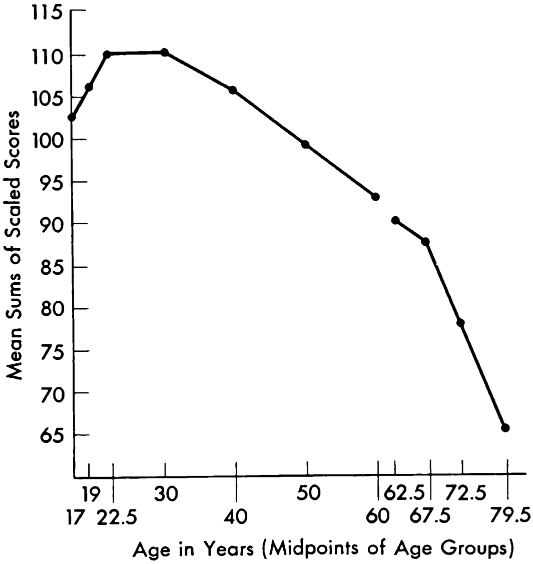


FIG. 55. Decline in WAIS Scaled Scores with Age.
(From Doppelt & Wallace, 1955, p. 323.)

group. Thus, if he shows the same decline in performance with age as the normative sample, his IQ will remain constant. The underlying assumption is that it is "normal" for an individual's tested ability to decline with age beyond the 30's.

Two facts from the Wechsler standardization data are relevant to the interpretation of age changes. First, since the standardization sample is a *normative* sample, it should reflect existing population characteristics at each age level (see Anastasi, 1956). Because of the rising educational level of the general population, older groups at any one point in time will have received less education than younger groups. This educational difference is clearly reflected in the WAIS standardization sample, in which the maximum years of schooling are found at the 20–34 year levels and educational level drops consistently in the older groups. These age differences in amount of education are inevitable if the standardization sample is to be truly representative of the population of the country at the time the norms were established. Nevertheless, the educational differences complicate the interpretation of the observed score decrements. The older groups in the standardization sample may have performed more poorly on the test, not because they were growing old, but because they had received less education than the younger groups.

A second pertinent fact emerges from a comparison of the WAIS with the Wechsler-Bellevue, which was standardized approximately 15 years earlier. In the Wechsler-Bellevue standardization sample, improvement in score ceased at an earlier age and the decline set in earlier than in the WAIS sample. An examination of the educational distributions of the two samples reveals that the changes in the age curves parallel the educational changes that have occurred in the general population during the intervening period. Persons in the WAIS standardization sample had received more education on the average than persons in the corresponding age in the Wechsler-Bellevue sample, since the latter were educated 15 years earlier.

The results obtained with the standardization samples of the Wechsler scales are typical of the findings of all cross-sectional studies of adult intelligence. Cross-sectional comparisons, in which persons of different ages are examined at the same time, are likely to show an apparent age decrement because cultural changes are confounded with the effects of aging. Amount of formal education is only one of many variables in which age groups may differ. Other cultural changes have occurred in our society during the past half century which make the experiential backgrounds of 20-year-olds and 70-year-olds quite dissimilar. Certainly changes in communication media, such as radio and television, and in transportation facilities have greatly increased the range of information available to the developing individual. Improvements in nutrition and medical care would also indirectly influence behavior development.

Longitudinal studies, based on retests of the same persons over periods of 5 to 40 years, have generally revealed the opposite trend, the scores tending to improve with age. Several of these longitudinal investigations have been conducted with intellectually superior groups, such as college graduates or individuals initially chosen because of high IQ's (Bayley & Oden, 1955; Burns, 1966; Campbell, 1965; Nisbet, 1957; Owens, 1953, 1966). For this reason, some writers have argued that the findings may be restricted to persons in the higher intellectual or educational levels and do not apply to the general population. However, similar results have been obtained in other longitudinal studies with normals (Charles & James, 1964; Eisdorfer, 1963; Schaie & Labouvie, 1974; Tuddenham, Blumenkrantz, & Wilkin, 1968), as well as with mentally retarded adults outside of institutions (Baller, Charles, & Miller, 1967; Bell & Zubek, 1960; Charles, 1953).

Neither cross-sectional nor longitudinal studies alone can provide a conclusive interpretation of observed age changes. Several excellent analyses of the methodological difficulties inherent in each approach, together with the required experimental designs, have been published (Baltes, 1968; Buss, 1973; Damon, 1965; Goulet & Baltes, 1970; Kuhlen, 1963; Nesselroade & Reese, 1973; Schaie, 1965). Basically, what is needed in order to tease out the effect of cultural changes is a combination of cross-sectional and longitudinal approaches. On the one hand, age differences in educational level may produce a spurious age decrement in test performance in cross-sectional studies. On the other hand, as the individual grows older, he is himself exposed to cultural changes that may improve his performance on intelligence tests.

A few studies provide data that permit at least a partial analysis of the contributing factors. Owens (1966) in his 40-year retest of Iowa State University students and D. P. Campbell (1965) in his 25-year retest of University of Minnesota students also tested *present* freshmen in the respective colleges. Thus, multiple comparisons could be made between the two groups tested at the same age 25 or 40 years apart, and the performance of a single group tested before and after the same time intervals. In both studies, the initial group improved over its own earlier performance, but performed about on a par with the younger group tested at the later date. Such findings suggest that it is cultural changes and other experiential factors, rather than age per se, that produce both the rises and declines in scores obtained with the more limited experimental designs.

Another, particularly well-designed study utilizing this combined approach was conducted on a more nearly representative sample of the adult population (Schaie & Strother, 1968). A specially selected battery of tests[2] was administered to a stratified-random sample of 500 persons.

---

[2] The SRA Primary Mental Abilities Test and Schaie's Test of Behavioral Rigidity.

The population from which this sample was drawn consisted of approximately 18,000 members of a prepaid medical plan, whose membership was fairly representative of the census figures for a large metropolitan area. The sample included 25 men and 25 women at each five-year age interval from 20 to 70. Seven years later, all the original subjects who could be located were contacted and 302 of them were given the same tests again. This subsample was shown to be closely comparable to the original group in age, sex ratio, and socioeconomic level.

The design of this study permits two types of comparisons: (1) a cross-sectional comparison among different age groups from 20 to 70 tested at the same time, and (2) a longitudinal comparison within the same individuals, initially tested at ages ranging from 20 to 70 and retested after seven years. The results of the cross-sectional comparisons showed significant intergeneration differences on all tests. In other words, those born and reared more recently performed better than those born and reared at an earlier time period. Longitudinal comparisons, on the other hand, showed a tendency for mean scores either to rise or remain unchanged when individuals were retested. The one major exception occurred in two highly speeded tests, in which performance was significantly poorer after the seven-year interval.

The contrast between the results of the cross-sectional and longitudinal approaches is illustrated in Figure 56, showing the trends obtained with four of the tests.[3] Similar results were found in a second seven-year retest of 161 of the original participants (Schaie & Labouvie-Vief, 1974). Further corroboration was provided by a still different approach involving the testing of three independent age-stratified samples drawn from the same population in 1956, 1963, and 1970 (Schaie, Labouvie, & Buech, 1973).

In general, the results of the better designed studies of adult intelligence strongly suggest that the ability decrements formerly attributed to aging are actually intergeneration or intercohort differences, probably associated with cultural changes in our society. Genuine ability decrements are not likely to be manifested until well over the age of 60. Moreover, any generalization, whether pertaining to age decrement or cohort differences, must be qualified by a recognition of the wide individual variability found in all situations. Individual differences within any one age level are much greater than the average difference between age levels. As a result, the distributions of scores obtained by persons of different ages overlap extensively. This simply means that large numbers of older persons can be found whose performance equals that of younger

---

[3] These are the tests that most closely resemble intelligence tests in their content. Of these four tests, only Reasoning showed a barely significant ($p < .05$) relation to age in the longitudinal study. The magnitude of the decline, however, is much smaller than in the cross-sectional comparison.

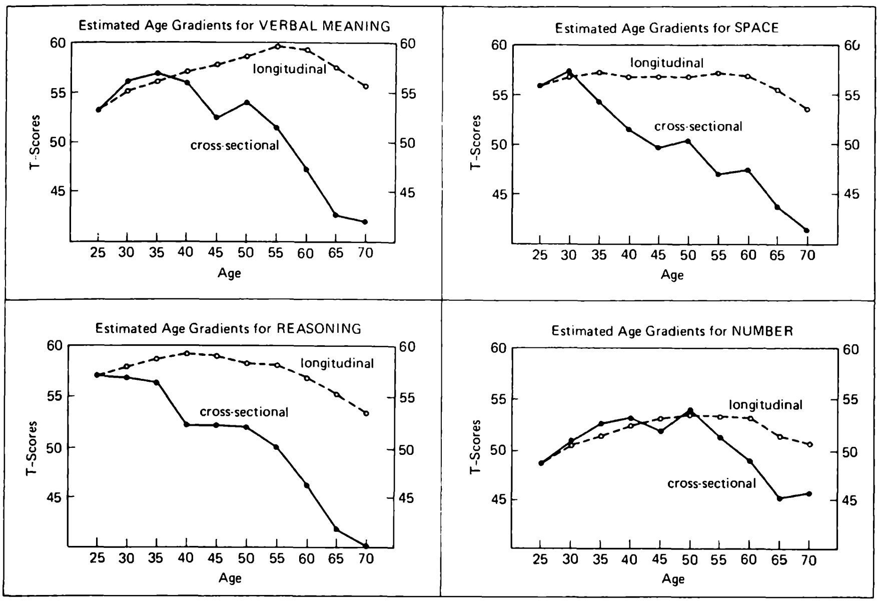


Fig. 56. Differences in Adult Intelligence as Assessed by Cross-Sectional and Longitudinal Studies.

(From Schaie & Strother, 1968, pp. 675, 676. Copyright 1968 by the American Psychological Association. Reprinted by permission.)

persons. Moreover, the best performers within the older groups excel the poorest performers within the younger groups. Nor is such overlapping limited to adjacent age levels; the ranges of performance still overlap when extreme groups are compared. Thus some 80-year-olds will do better than some 20-year-olds.

What is even more relevant to the present topic, however, is that the *changes* that occur with aging vary with the individual. Thus between the ages of 50 and 60, for example, some persons may show a decrease, some no appreciable change, and some an increase in test performance. The amount of change, whether it be a drop or a rise, will also vary widely among individuals. Moreover, intensive studies of persons of advanced age, extending into the seventh, eighth, and ninth decades of life, indicate that intellectual functioning is more closely related to the individual's health status than to his chronological age (Birren, 1968; Palmore, 1970).

NATURE OF ADULT INTELLIGENCE. Within the life span, testing has been oriented chiefly toward the schoolchild and college student. At these levels, the test constructor can draw on the large common pool of experiences that have been organized into academic curricula. Most intelligence tests measure how well the individual has acquired the intellectual skills taught in our schools; and they can in turn predict how well he is prepared for the next level in the educational hierarchy. Tests for adults, including the Wechsler scales, draw largely on this identifiable common fund of experience. As the individual grows older and his own formal educational experiences recede farther into the past, this fund of common experience may become increasingly less appropriate to assess his intellectual functioning. Adult occupations are more diversified than childhood schooling. The cumulative experiences of adulthood may thus stimulate a differential development of abilities in different persons.

Because intelligence tests are closely linked to academic abilities, it is not surprising to find that longitudinal studies of adults show larger age increments in score among those individuals who have continued their education longer (D. P. Campbell, 1965; Härnqvist, 1968; Husén, 1951; Lorge, 1945; Owens, 1953). Similarly, persons whose occupations are more "academic" in content, calling into play verbal and numerical abilities, are likely to maintain their performance level or show improvement in intelligence test scores over the years, while those engaged in occupations emphasizing mechanical activities or interpersonal relations may show a loss. Some suggestive data in support of this hypothesis are reported by Williams (1960), who compared the performance of 100 persons, ranging in age from 65 to over 90, on a series of verbal and nonverbal tests. Rather striking correspondences were found between the individual's occupation and his relative performance on the two types of tasks. Longitudinal investigations of adults have also found suggestive relationships between total IQ changes and certain biographical inventory items (Charles & James, 1964; Owens, 1966).

Each time and place fosters the development of skills appropriate to its characteristic demands. Within the life span, these demands differ for the infant, the schoolchild, the adult in different occupations, and the retired septuagenarian. An interesting demonstration of the implications of this fact for intelligence testing was provided by Demming and Pressey (1957). These investigators began with a task analysis of typical adult functions, conducted through informal surveys of reading matter and of reported daily activities and problems. On this basis, they prepared preliminary forms of some 20 tests "indigenous" to the older years. The tests emphasized practical information, judgment, and social perception. Results with three of these tests, administered together with standard verbal and nonverbal tests to samples of different ages, showed that the older persons excelled the younger on the new tests while the reverse relationship held for the traditional tests. All these types of research suggest that whether intelligence test scores rise or decline with increasing age in adulthood depends largely on what experiences the individual undergoes during those years and on the relationship between these experiences and the functions covered by the tests.

PROBLEMS IN CROSS-CULTURAL TESTING

The use of tests with persons of diverse cultural backgrounds has already been considered from various angles in earlier parts of this book. Chapter 3 was concerned with the social and ethical implications of such testing, particularly with reference to minority groups within a broader national culture. Technical problems pertaining to test bias and to item-group interaction were analyzed in Chapters 7 and 8. And in Chapter 10 we examined typical tests designed for various transcultural applications. In this section, we shall present some basic theoretical issues about the role of culture in behavior, with special reference to the interpretation of intelligence test scores.

LEVELS OF CULTURAL DIFFERENTIALS. Cultural differences may operate in many ways to bring about group differences in behavior. The level at which cultural influences are manifested varies along a continuum extending from superficial and temporary effects to those which are basic, permanent, and far-reaching. From both a theoretical and a practical standpoint, it is important to inquire at what level of this continuum any observed behavioral difference falls. At one extreme we find cultural differences that may affect only responses on a particular test and thus reduce its validity for certain groups. There are undoubtedly test items that have no diagnostic value when applied to persons from certain cultures because of lack of familiarity with specific objects or other relatively trivial experiential differences.

Most cultural factors that affect test responses, however, are also likely to influence the broader behavior domain that the test is designed to sample. In an English-speaking culture, for example, inadequate mastery of English may handicap a child not only on an intelligence test but also in his school work, contact with associates, play activities, and other situations of daily life. Such a condition would thus interfere with the child's subsequent intellectual and emotional development and would have practical implications that extend far beyond immediate test performance. At the same time, deficiencies of this sort can be remedied without much difficulty. Suitable language training *can* bring the individual up to an effective functioning level within a relatively short period.

The language an individual has been taught to speak was chosen in the above example because it provides an extreme and obvious illustration of several points: (1) it is clearly not a hereditary condition; (2) it can be altered; (3) it can seriously affect performance on a test administered in a different language; (4) it will similarly affect the individual's educational, vocational, and social activities in a culture that uses an unfamiliar language. Many other examples can be cited from the middle range of the continuum of cultural differentials. Some are cognitive differentials, such as reading disability or ineffective strategies for solving abstract problems; others are attitudinal or motivational, such as lack of interest in intellectual activities, hostility toward authority figures, low achievement drive, or poor self-concept. All such conditions can be ameliorated by a variety of means, ranging from functional literacy training to vocational counseling and psychotherapy. All are likely to affect both test performance and the daily life activities of the child and adult.

As we move along the continuum of cultural differentials, we must recognize that the longer an environmental condition has operated in the individual's lifetime, the more difficult it becomes to reverse its effects. Conditions that are environmentally determined are not necessarily remediable. Adverse experiential factors operating over many years may produce intellectual or emotional damage that can no longer be eliminated by the time intervention occurs. It is also important to bear in mind, however, that the permanence or irremediability of a psychological condition is no proof of hereditary origin.

An example of cultural differentials that may produce permanent effects on individual behavior is provided by research on complications of pregnancy and parturition (Knobloch & Pasamanick, 1966; Pasamanick & Knobloch, 1966). In a series of studies on large samples of blacks and whites, prenatal and perinatal disorders were found to be significantly related to mental retardation and behavior disorders in the offspring. An important source of such irregularities in the process of childbearing and birth is to be found in deficiencies of maternal nutrition and other conditions associated with low socioeconomic status. Analysis of the data revealed a much higher frequency of all such medical complications in lower than in higher socioeconomic levels, and a higher frequency among blacks than among whites. Here then is an example of cultural differentials producing organic disorders that in turn may lead to behavioral deficiencies. The effects of this type of cultural differential cannot be completely reversed within the lifetime of the individual, but require more than one generation for their elimination. Again it needs to be emphasized, however, that such a situation does not imply hereditary defect, nor does it provide any justification for failure to improve the environmental conditions that brought it about.

CULTURAL DIFFERENCES AND CULTURAL HANDICAP. When psychologists began to develop instruments for cross-cultural testing in the first quarter of this century, they hoped it would be at least theoretically possible to measure hereditary intellectual potential independently of the impact of cultural experiences. The individual's behavior was thought to be overlaid with a sort of cultural veneer whose penetration became the objective of what were then called "culture-free" tests. Subsequent developments in genetics and psychology have demonstrated the fallacy of this concept. We now recognize that hereditary and environmental factors interact at all stages in the organism's development and that their effects are inextricably intertwined in the resulting behavior. For man, culture permeates nearly all environmental contacts. Since all behavior is thus affected by the cultural milieu in which the individual is reared and since psychological tests are but samples of behavior, cultural influences will and should be reflected in test performance. It is therefore futile to try to devise a test that is *free* from cultural influences. The present objective in cross-cultural testing is rather to construct tests that presuppose only experiences that are *common* to different cultures. For this reason, such terms as "culture-common," "culture-fair," and "cross-cultural" have replaced the earlier "culture-free" label.

No single test can be universally applicable or equally "fair" to all cultures. There are as many varieties of culture-fair tests as there are parameters in which cultures differ. A nonreading test may be culture-fair in one situation, a nonlanguage test in another, a performance test in a third, and a translated adaptation of a verbal test in a fourth. The varieties of available cross-cultural tests are not interchangeable but are useful in different types of cross-cultural comparisons.

It is unlikely, moreover, that any test can be equally "fair" to more than one cultural group, especially if the cultures are quite dissimilar. While reducing cultural differentials in test performance, cross-cultural tests cannot completely eliminate such differentials. Every test tends to favor persons from the culture in which it was developed. The mere use of paper and pencil or the presentation of abstract tasks having no immediate practical significance will favor some cultural groups and handicap others. Emotional and motivational factors likewise influence test performance. Among the many relevant conditions differing from culture to culture may be mentioned the intrinsic interest of the test content, rapport with the examiner, drive to do well on a test, desire to excel others, and past habits of solving problems individually or cooperatively. In testing children of low socioeconomic level, several investigators have found that the examinees rush through the test, marking answers almost at random and finishing before time is called (Eells et al., 1951). The same reaction has been observed among Puerto Rican schoolchildren tested in New

York City and in Hawaii (Anastasi & Cordova, 1953; S. Smith, 1942). Such a reaction may reflect a combination of lack of interest in the relatively abstract test content and expectation of low achievement on tasks resembling school work. By hurrying through the test, the child shortens the period of discomfort.

Each culture and subculture encourages and fosters certain abilities and ways of behaving, and discourages or suppresses others. It is therefore to be expected that, on tests developed within the majority American culture, for example, persons reared in that culture will generally excel. If a test were constructed by the same procedures within a culture differing markedly from ours, Americans would probably appear deficient in terms of test norms. Data bearing on this type of cultural comparison are meager. What evidence is available, however, suggests that persons from our culture may be just as handicapped on tests prepared within other cultures as members of those cultures are on our tests (Anastasi, 1958a, pp. 566–568). Cultural differences become cultural handicaps when the individual moves out of the culture or subculture in which he was reared and endeavors to function, compete, or succeed within another culture. From a broader viewpoint, however, it is these very contacts and interchanges between cultures that stimulate the advancement of civilizations. Cultural isolation, while possibly more comfortable for individuals, leads to societal stagnation.

LANGUAGE IN TRANSCULTURAL TESTING. Most cross-cultural tests utilize nonverbal content in the hope of obtaining a more nearly culture-fair measure of the same intellectual functions measured by verbal intelligence tests. Both assumptions underlying this approach are questionable. First, it cannot be assumed that nonverbal tests measure the same functions as verbal tests, however similar they may appear. A spatial analogies test is not merely a nonverbal version of a verbal analogies test. Some of the early nonlanguage tests, such as the Army Beta, were heavily loaded with spatial visualization and perceptual abilities, which are quite unrelated to verbal and numerical abilities. Even in tests like the Progressive Matrices and other nonlanguage tests deliberately designed to tap reasoning and abstract conceptualization, factorial analyses have revealed a large contribution of nonverbal factors to the variance of test scores (e.g., Das, 1963).

From a different angle, there is a growing body of evidence suggesting that nonlanguage tests may be more culturally loaded than language tests. Investigations with a wide variety of cultural groups in many countries have found larger group differences in performance and other nonverbal tests than in verbal tests (Anastasi, 1961; Irvine, 1969a, 1969b; Jensen, 1968; Ortar, 1963, 1972; Vernon, 1965). In a provocative analysis of the problem, Ortar (1963, pp. 232–233) writes:

> On the basis of our results it appears that, both from the practical point of view and on theoretical grounds, the verbal tests and items are better suited as intercultural measuring instruments than any other kind. They must, of course, be translated and adapted, but this adaptation is infinitely easier and more reliable than the well-nigh impossible task of "translating" and adapting a performance test. The "language" of performance is the cultural perception, but its "words" and grammar and syntax are not even completely understood, let alone organized in national entities. We do not know how to "translate" a picture into the representational language of a different culture, but we are thoroughly familiar with the technique and requirements of translating verbal contents. . . . A concept that is non-existent in a certain language simply cannot be translated into this language, a factor which acts as a safeguard against mechanical use of a given instrument when adapting it for a different culture.

Among the examples cited by Ortar is the observation that, when presented with a picture of a head from which the mouth was missing, Oriental immigrant children in Israel said the body was missing. Unfamiliar with the convention of considering the drawing of a head as a complete picture, these children regarded the absence of the body as more important than the omission of a mere detail like the mouth. For a different reason, an item requiring that the names of the seasons be arranged in the proper sequence would be more appropriate in a cross-cultural test than would an item using pictures of the seasons. The seasons would not only look different in different countries for geographical reasons, but they would also probably be represented by means of conventionalized pictorial symbols which would be unfamiliar to persons from another culture.

The use of pictorial representation itself may be unsuitable in cultures unaccustomed to representative drawing. A two-dimensional reproduction of an object is not a perfect replica of the original; it simply presents certain cues which, as a result of past experience, lead to the perception of the object. If the cues are highly reduced, as in a simplified or schematic drawing, and if the necessary past experience is absent, the correct perception may not follow. There is now a considerable body of empirical data indicating marked differences in the perception of pictures by persons in different cultures (Miller, 1973; Segall, Campbell, & Herskovits, 1966).

From still another angle, nonverbal tests frequently require relatively abstract thinking processes and analytic cognitive styles characteristic of middle-class Western cultures (R. A. Cohen, 1969). Persons reared in other cultural contexts may be much less accustomed to such problem-solving approaches.

It should be added that nonverbal tests have fared no better in the testing of minority groups and persons of low socioeconomic status within the United States. On the WISC, for instance, black children usually find the Performance tests as difficult as or more difficult than the Verbal tests; this pattern is also characteristic of children from low socioeconomic levels (Caldwell & Smith, 1968; Cole & Hunter, 1971; Goffeney, Henderson, & Butler, 1971; Hughes & Lessler, 1965; Teahan & Drews, 1962). The same groups tend to do better on the Stanford-Binet than on either Raven's Progressive Matrices (Higgins & Sivers, 1958) or Cattell's Culture Fair Intelligence Test (Willard, 1968).

There is of course no procedural difficulty in administering a verbal test across cultures speaking a common language. When language differences necessitate a translation of the test, problems arise regarding the comparability of norms and the equivalence of scores. It should also be noted that a simple translation would rarely suffice. Some adaptation and revision of content is generally required. Of interest in this connection is the procedure developed in equating the scales of the CEEB Scholastic Aptitude Test (SAT) and the Puebra de Aptitud Académica (PAA)—a Spanish version of the SAT (Angoff & Modu, 1973).

The PAA was originally developed for local use in Puerto Rico, but has subsequently been considered by continental American universities as a possible aid in the admission of Spanish-speaking students. With such objectives in mind, an exploratory project in scale equating was undertaken. This project provides a demonstration of an ingenious method applicable to other situations requiring testing in multiple languages.

Essentially, the procedure consists of two steps. The first step is the selection of a common set of anchor items judged to be equally appropriate for both groups of students. These items are administered in English to the English-speaking students and in Spanish to the Spanish-speaking students. The performance of these groups provides the data for measuring the difficulty level (Δ value) and discriminative power (r_{bis} with total test score) of each item. For the final set of anchor items, any items showing appreciable item-group interaction are discarded (see Ch. 8). These would be the "biased" items that are likely to have a psychologically different meaning for the two groups. The final anchor items are those having approximately the same *relative* difficulty for the English-speaking and the Spanish-speaking samples, in addition to meeting the specifications for difficulty level and discriminative power.

The second step is to include these anchor items in a regular administration of the SAT and the PAA and to use the scores on the anchor items as a basis for converting all test scores to a single scale. This step involves the same equating procedure regularly followed in converting scores on successive forms of the SAT to a uniform scale.

NATURE OF INTELLIGENCE

MEANING OF AN IQ. For the general public, the IQ is not identified with a particular type of score on a particular test, but is often a shorthand designation for intelligence. So prevalent has this usage become, that it cannot be merely ignored or deplored as a popular misconception. To be sure, when considering the numerical value of a given IQ, we should always specify the test from which it was derived. Different intelligence tests that yield an IQ do in fact differ in content and in other ways that affect the interpretation of their scores. Some of these differences among tests sharing the common label of "intelligence test" were apparent in the examples considered in the preceding chapters. Nonetheless, there is a need to reexamine the general connotations of the construct "intelligence," as symbolized by the IQ. It might be added that the prevalent conception of intelligence has been shaped to a considerable degree by the characteristics of the Stanford-Binet scale, which for many years provided the only instrument for the intensive measurement of intelligence and which was often used as a criterion for validating new tests.

First, intelligence should be regarded as a descriptive rather than an explanatory concept. An IQ is an expression of an individual's ability level at a given point in time, in relation to his age norms. No intelligence test can indicate the reasons for his performance. To attribute inadequate performance on a test or in everyday-life activities to "inadequate intelligence" is a tautology and in no way advances our understanding of the individual's handicap. In fact, it may serve to halt efforts to explore the causes of the handicap in the individual's history.

Intelligence tests, as well as any other kind of tests, should be used not to label an individual but to help in understanding him. To bring a person to his maximum functioning level we need to start where he is at the time; we need to assess his strengths and weaknesses and plan accordingly. If a reading test indicates that a child is retarded in reading, we do not label him as a nonreader and stop; nor do we give him a nonverbal test to conceal his handicap. Instead we concentrate on teaching him to read.

An important goal of contemporary testing, moreover, is to contribute to self-understanding and personal development. The information provided by tests is being used increasingly to assist individuals in educational and vocational planning and in making decisions about their own lives. The attention being given to effective ways of communicating test results to the individual attests to the growing recognition of this application of testing.

A second major point to bear in mind is that intelligence is not a single, unitary ability, but a composite of several functions. The term is commonly used to cover that combination of abilities required for survival and advancement within a particular culture. It follows that the specific abilities included in this composite, as well as their relative weights, will vary with time and place. In different cultures and at different historical periods within the same culture, the qualifications for successful achievement will differ. The changing composition of intelligence can also be recognized within the life of the individual, from infancy to adulthood. An individual's relative ability will tend to increase with age in those functions whose value is emphasized by his culture or subculture; and his relative ability will tend to decrease in those functions whose value is deemphasized (see, e.g., Levinson, 1959, 1961).

Typical intelligence tests designed for use in our culture with school-age children or adults measure largely verbal abilities; to a lesser degree, they also cover abilities to deal with numerical and other abstract symbols. These are the abilities that predominate in school learning. Most intelligence tests can therefore be regarded as measures of scholastic aptitude. The IQ is both a reflection of prior educational achievement and a predictor of subsequent educational performance. Because the functions taught in the educational system are of basic importance in our culture, the IQ is also an effective predictor of performance in many occupations and other activities of adult life.

On the other hand, there are many other important functions that intelligence tests have never undertaken to measure. Mechanical, motor, musical, and artistic aptitudes are obvious examples. Motivational, emotional, and attitudinal variables are important determiners of achievement in all areas. Current creativity research is identifying both cognitive and personality variables that are associated with creative productivity. All this implies, of course, that both individual and institutional decisions should be based on as much relevant data as can reasonably be gathered. To base decisions on tests alone, and especially on one or two tests alone, is clearly a misuse of tests. Decisions must be made by persons. Tests represent one source of data utilized in making decisions; they are not themselves decision-making instruments.

HERITABILITY AND MODIFIABILITY. Much confusion and controversy have resulted from the application of heritability estimates to intelligence test scores. A well-known example is an article by Jensen (1969), which has engendered great furor and led to many heated arguments. Although there are several aspects to this controversy and the issues are complicated, a major substantive source of controversy pertains to the interpretation of heritability estimates. Specifically, a heritability index shows the proportional contribution of genetic or hereditary factors to the total variance of a particular trait in a given population under existing conditions. For example, the statement that the heritability of Stanford-Binet IQ among urban American high school students is .70 would mean that 70 percent of the variance found in these scores is attributable to hereditary differences and 30 percent is attributable to environment.

Heritability indexes have been computed by various formulas (see, e.g., Jensen, 1969; Loehlin, Lindzey, & Spuhler, 1975), but their basic data are measures of familial resemblance in the trait under consideration. A frequent procedure is to utilize intelligence test correlations of monozygotic (identical) and dizygotic (fraternal) twins. Correlations between monozygotic twins reared together and between monozygotic twins reared apart in foster homes have also been used.

Several points should be noted in interpreting heritability estimates. First, the empirical data on familial resemblances are subject to some distortion because of the unassessed contributions of environmental factors. For instance, there is evidence that monozygotic twins share a more closely similar environment than do dizygotic twins (Anastasi, 1958a, pp. 287–288; Koch, 1966). Another difficulty is that twin pairs reared apart are not assigned at random to different foster homes, as they would be in an ideal experiment; it is well known that foster home placements are selective with regard to characteristics of the child and the foster family. Hence the foster home environments of the twins within each pair are likely to show sufficient resemblance to account for some of the correlation between their test scores. There is also evidence that twin data regarding heritability may not be generalizable to the population at large because of the greater susceptibility of twins to prenatal trauma leading to severe mental retardation. The inclusion of such severely retarded cases in a sample may greatly increase the twin correlation in intelligence test scores (Nichols & Broman, 1974).

Apart from questionable data, heritability indexes have other intrinsic limitations (see, Anastasi, 1971; Hebb, 1970). It is noteworthy that in the early part of the previously cited article, Jensen (1969, pp. 33–46) clearly lists these limitations among others. First, the concept of heritability is applicable to populations, not individuals. For example, in trying to establish the etiology of a particular child's mental retardation, the heritability index would be of no help. Regardless of the size of the heritability index in the population, the child's mental retardation could have resulted from a defective gene (as in phenylketonuria or PKU), from prenatal brain damage, or from extreme experiential deprivation.

Second, heritability indexes refer to the population on which they were found at the time. Any change in either hereditary or environmental conditions would alter the heritability index. For instance, an increase in inbreeding, as on an isolated island, would reduce the variance attributable to heredity and hence lower the heritability index; increasing environmental homogeneity, on the other hand, would reduce the variance attributable to environment and hence raise the heritability index. Furthermore, a heritability index computed within one population is not applicable to an analysis of the differences in test performance between two populations, such as different ethnic groups.

Third, heritability does not indicate the degree of modifiability of a trait. Even if the heritability index of a trait in a given population is 100 percent, it does not follow that the contribution of environment to that trait is unimportant. An extreme example may help to clarify the point. Suppose in a hypothetical adult community everyone has the identical diet. All receive the same food in identical quantities. In this population, the contribution of food to the total variance of health and physical condition would be zero, since food variance accounts for none of the individual differences in health and physique. Nevertheless, if the food supply were suddenly cut off, the entire community would die of starvation. Conversely, improving the quality of the diet could well result in a general rise in the health of the community.

Regardless of the magnitude of heritability indexes found for IQ's in various populations, one empirical fact is well established: the IQ is not fixed and unchanging; and it is amenable to modification by environmental interventions. Some evidence for this conclusion was examined earlier in this chapter, in connection with longitudinal studies. There has been some progress in identifying characteristics of accelerating and decelerating environments. Rises and drops in IQ may also result from both fortuitous environmental changes occurring in a child's life and planned environmental interventions (see Anastasi, 1958a). Major changes in family structure, sharp rises or drops in family income level, adoption into a foster home, or participation in a preschool program may produce conspicuous increases or decreases in IQ.

Prediction of both criterion performance and subsequent test score can be improved by taking into account salient features of the environment in which each child has lived during the intervening period. There is some suggestive evidence that correlations with subsequent intelligence test score or academic achievement can be substantially raised when environmental variables are included along with initial test scores as predictors (Bloom, 1964, Ch. 6). For example, in a group of 40 pupils, the correlation between reading comprehension at grade 2 and at grade 8 rose from .52 to .72 when father's occupation was added as a rough index of the cultural level of the home (p. 119). Again, a reanalysis of intelligence test data from the Harvard Growth Study showed that the correlation between intelligence test scores at ages 7 and 16 could be raised from .58 to .92, when parents' educational level was included in the multiple correlation (p. 80).

Relevant data are also provided by longitudinal studies of populations, in which comparable samples of the same population are tested many years apart (see Anastasi, 1958a, pp. 209–211). When cultural conditions have improved over the interval, a significant rise in the mean intelligence test performance of the population has generally been found. Such findings are illustrated by a comparison of the test scores of American soldiers examined during World Wars I and II, spanning an interval of 25 years (Tuddenham, 1948). For this purpose, a representative sample of enlisted men in the World War II army was given both the AGCT and a revision of the Army Alpha of World War I. The distribution of this group on the AGCT paralleled closely that of the entire army. With the data from this sample as a bridge, it was possible to estimate that the median performance of the World War II army fell at the 83rd percentile of the World War I army. In other words, 83 percent of the World War I population fell below the median score of the World War II population. It is noteworthy that the average amount of education in the World War II population was 10 years, as compared with 8 years during World War I. This increase in amount of education, accompanied by improvements in communication, transportation, and other cultural changes that increased the individual's range of experience, was reflected in improved test performance.

A similar investigation on a smaller scale was conducted with schoolchildren living in certain mountainous counties in Tennessee (Wheeler, 1942). Group intelligence tests were administered in 1940 to over 3,000 children in 40 rural schools. The results were compared with those obtained with children in the same areas and largely from the same families, who had been similarly tested in 1930. During the intervening 10-year period, the economic, cultural, and educational status of these counties had improved conspicuously. Paralleling such environmental changes, a rise in IQ was found at all ages and all grades. The median IQ was 82 in the 1930 sample and 93 in the 1940 sample.

From another angle, the very composition of intelligence may alter as a result of the individual's experiences. The individual's pattern of abilities will tend to change with age as his environment fosters and encourages the development of some aptitudes and deemphasizes other aptitudes. Moreover, factor-analytic research has demonstrated that experiential differences may influence not only the level of performance reached in different abilities, but also the way in which intelligence becomes differentiated into identifiable traits.[4] There is empirical evidence that the number and nature of traits or abilities may change over time and may differ among cultures or subcultures (Anastasi, 1970).

---

[4] Further discussion of this point will be found in Chapter 13.

An individual's intelligence at any one point in time is the end product of a vast and complex sequence of interactions between hereditary and environmental factors. At any stage in this causal chain, there is opportunity for interaction with new factors; and because each interaction in turn determines the direction of subsequent interactions, there is an ever-widening network of possible outcomes. The connection between the genes an individual inherits and any of his behavioral characteristics is thus highly indirect and devious (see Anastasi, 1958b, 1973; Hebb, 1953).

INTELLIGENCE AND PERSONALITY. Although it is customary and convenient to classify tests into separate categories, it should be recognized that all such distinctions are superficial. In interpreting test scores, personality and aptitudes cannot be kept apart. An individual's performance on an aptitude test, as well as his performance in school, on the job, or in any other context, is influenced by his achievement drive, his persistence, his value system, his freedom from handicapping emotional problems, and other characteristics traditionally classified under the heading of "personality."

Even more important is the cumulative effect of personality characteristics on the direction and extent of the individual's intellectual development. Some of the evidence for this effect, collected through longitudinal studies of children and adults, was summarized earlier in this chapter. Other investigations on groups ranging from preschool children to college students have been surveyed by Dreger (1968). Although some of the research on young children utilized a longitudinal approach, data from older subjects were gathered almost exclusively through concurrent correlations of personality test scores with intelligence test scores and indices of academic achievement. The data assembled by Dreger indicate the importance of considering appropriate personality variables as an aid in understanding an individual's intelligence test performance and in predicting his academic achievement.

It would thus seem that prediction of a child's subsequent intellectual development could be improved by combining information about his emotional and motivational characteristics with his scores on ability tests. A word should be added, however, regarding the assessment of "motivation." In the practical evaluation of schoolchildren, college students, job applicants, and other categories of persons, psychologists are often asked for a measure of the individual's "motivation." When thus worded, this is a meaningless request, since motivation is specific. What is needed is an indication of the individual's value system and the intensity with whch he will strive toward specific goals. The strength of these specific motives will interact with situational factors, as well as with aptitudes, to determine the individual's actual performance in given situations.

The relation between personality and intellect is reciprocal. Not only do personality characteristics affect intellectual development, but intellectual level also affects personality development. Suggestive data in support of this relation are provided in a study by Plant and Minium (1967). Drawing upon the data gathered in five available longitudinal investigations of college-bound young adults, the authors selected the upper and lower 25 percent of each sample in terms of intelligence test scores. These contrasted groups were then compared on a series of personality tests that had been administered to one or more of the samples. The personality tests included measures of attitudes, values, motivation, and interpersonal and other noncognitive traits. The results of this analysis revealed a strong tendency for the high-aptitude groups to undergo substantially more "psychologically positive" personality changes than did the low-aptitude groups.

The success the individual attains in the development and use of his aptitudes is bound to influence his emotional adjustment, interpersonal relations, and self-concept. In the self-concept we can see most clearly the mutual influence of aptitudes and personality traits. The child's achievement in school, on the playground, and in other situations helps to shape his self-concept; and his self-concept at any given stage influences his subsequent performance, in a continuing spiral. In this regard, the self-concept operates as a sort of private self-fulfilling prophecy.

At a more basic theoretical level, K. J. Hayes (1962) has proposed a broadly oriented hypothesis concerning the relationship of drives and intellect. Regarding intelligence as a collection of learned abilities, Hayes maintains that the individual's motivational makeup influences the kind and amount of learning that occurs. Specifically, it is the strength of the "experience-producing drives" that affects intellectual development. These drives are illustrated by exploratory and manipulatory activities, curiosity, play, the babbling of infants, and other intrinsically motivated behavior. Citing chiefly research on animal behavior, Hayes argues that these experience-producing drives are genetically determined and represent the only hereditary basis of individual differences in intelligence. It might be added that the hereditary or environmental basis of the experience-producing drives need not alter the conceptualization of their role in intellectual development. These two parts of the theory may be considered independently.

Whatever the origin of the experience-producing drives, the individual's experience is regarded as a joint function of the strength of these drives and the environment in which they operate. The cumulative effect of these experiences in turn determines the individual's intellectual level at any given time. This is a provocative hypothesis, through which Hayes integrates a considerable body of data from many types of research on both human and animal behavior.

On the basis of 25 years of research on achievement motivation, J. W. Atkinson (1974, 1976) and his co-workers have formulated a comprehensive schema representing the interrelationships of abilities, motivation, and environmental variables. The approach is dynamic in that it implies systematic change rather than constancy in the individual's lifetime. It also incorporates the reciprocal effects of aptitudes and motivational variables; and it emphasizes the contribution of motivational variance to test performance. To illustrate the application of this conceptual schema, computer simulations were employed, showing how ability and motivation can jointly influence both intelligence test performance and cumulative achievement. Some supporting empirical data are also cited regarding the high school grade-point average of boys as predicted from earlier intelligence test scores and a measure of achievement motivation (Atkinson, O'Malley & Lens, 1976).

A diagram of Atkinson's conceptual schema is reproduced in Figure 57. Beginning at the left, this figure shows the combined operation of heredity and past formative environment in the development of both cognitive and noncognitive aspects of the total personality. The present

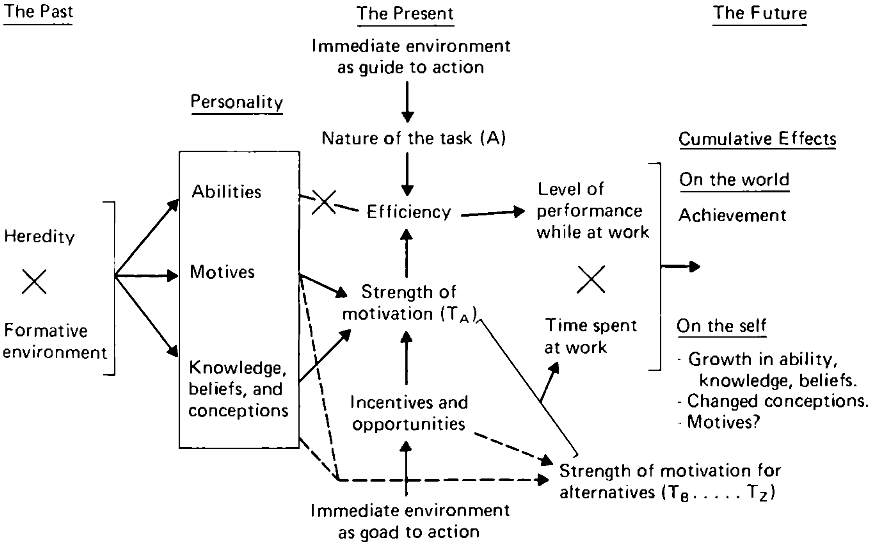

The Past
The Present
The Future
Immediate environment as guide to action
Personality
Nature of the task (A)
Cumulative Effects
Abilities
Efficiency
Level of performance while at work
On the world
Achievement
Heredity
×
Formative environment
Motives
Strength of motivation (T_A)
×
Time spent at work
On the self
Knowledge, beliefs, and conceptions
Incentives and opportunities
- Growth in ability, knowledge, beliefs.
- Changed conceptions.
- Motives?
Strength of motivation for alternatives (T_B . . . . . T_Z)
Immediate environment as goad to action

Fig. 57. Schema Illustrating Interaction of Cognitive and Noncognitive Factors in Cumulative Achievement and Individual Development.

(From Atkinson, O'Malley, & Lens, 1976, Ch. 8. Reproduced by permission of Academic Press.)

environment provides the immediate task and contributes to motivational strength for this task relative to the competing motivational strengths of alternative actions. Motivation affects both the efficiency with which the task is performed and the time spent on it (e.g., studying, carrying out a job-related activity). Efficiency reflects the relationship between nature of task and current motivation. Level of performance results from the individual's relevant abilities (e.g., as assessed by test scores) and the efficiency with which he applies those abilities to the current task. The final achievement or product shows the combined effects of level of performance while at work and time spent at work. Another and highly important consequent of level of performance × time spent at work is the lasting cumulative effects of this activity or experience on the individual's own cognitive and noncognitive development. This last step represents a feedback loop to the individual's personality, whose effects are likely to be reflected in his future test scores.

Many implications of this approach for the interpretation of test scores are spelled out in the original sources cited above, which should be examined for further details. The schema provides a promising orientation for the more effective utilization of tests and for the better understanding of the conditions underlying human achievement.

PART 4

Tests of Separate Abilities

CHAPTER 13

Measuring Multiple Aptitudes

ONE OF the distinguishing features of contemporary psychological testing is its *differential approach* to the measurement of ability. To be sure, the general intelligence tests discussed in Part 3 are still widely used in preliminary screening and in the clinical examination of extreme deviates. Since midcentury, however, there has been a rapid increase in the development and application of instruments that permit an analysis of performance with regard to different aspects of intelligence. Such instruments yield, not a single, global measure such as an IQ, but a set of scores in different aptitudes. They thus provide an intellectual *profile* showing the individual's characteristic strengths and weaknesses.

A number of events have contributed to the growing interest in differential aptitude testing. First, there has been an increasing recognition of intraindividual variation in performance on intelligence tests. Crude attempts to compare the individual's relative standing on different subtests or item groups antedated the development of multiple aptitude batteries by many years. As has been repeatedly pointed out, however, intelligence tests were not designed for this purpose. The subtests or item groups are often too unreliable to justify intraindividual comparisons. In the construction of intelligence tests, moreover, items or subtests are generally chosen to provide a unitary and internally consistent measure. In such a selection, an effort is therefore made to minimize, rather than maximize, intraindividual variation. Subtests or items that correlate very low with the rest of the scale would, in general, be excluded. Yet these are the very parts that would probably have been retained if the emphasis had been on the differentiation of abilities. Because of the way in which intelligence tests are constructed, it is unlikely that performance on these tests can be significantly differentiated into more than two categories, such as verbal and nonverbal or linguistic and quantitative.

The development of multiple aptitude batteries has been further stimulated by the gradual realization that so-called general intelligence tests are in fact less general than was originally supposed. It soon became apparent that many such tests were primarily measures of verbal comprehension. Certain areas, such as that of mechanical abilities, were usually untouched, except in some of the performance and nonlanguage scales. As these limitations of intelligence tests became evident, psychologists began to qualify the term "intelligence." Distinctions between "academic" and "practical" intelligence were suggested by some. Others spoke of "abstract," "mechanical" and "social" intelligence. Tests of "special aptitudes" were likewise designed to supplement the intelligence tests. But closer analysis showed that the intelligence tests themselves could be said to measure a certain combination of special aptitudes, such as verbal and numerical abilities, although the area covered by these tests was loosely and inconsistently defined.

A strong impetus to differential aptitude testing was also provided by the growing activities of psychologists in vocational counseling, as well as in the selection and classification of industrial and military personnel. The early development of specialized tests in clerical, mechanical, and other vocational areas is a reflection of such interests. The assembling of test batteries for the selection of applicants for admission to schools of medicine, law, engineering, dentistry, and other professional fields represents a similar development which has been in progress for many years. Moreover, a number of differential aptitude batteries, such as those prepared by the armed services and by the United States Employment Service, were the direct result of vocational selection or classification work.

Finally, the application of factor analysis to the study of trait organization provided the theoretical basis for the construction of multiple aptitude batteries. Through factor-analytic techniques, the different abilities loosely grouped under "intelligence" could be more systematically identified, sorted, and defined. Tests could then be selected so that each represented the best available measure of one of the traits or factors identified by factor analysis.

FACTOR ANALYSIS

THE FACTOR MATRIX. The principal object of factor analysis is to simplify the description of data by reducing the number of necessary variables, or dimensions. Thus, if we find that five factors are sufficient to account for all the common variance in a battery of 20 tests, we can for most purposes substitute 5 scores for the original 20 without sacrificing any essential information. The usual practice is to retain from among the original tests those providing the best measures of each of the factors.

All techniques of factor analysis begin with a complete table of intercorrelations among a set of tests. Such a table is known as a correlation matrix. Every factor analysis ends with a factor matrix, i.e., a table showing the weight or loading of each of the factors in each test. A hypothetical factor matrix involving only two factors is shown in Table 26. The factors are listed across the top and their weights in each of the 10 tests are given in the appropriate rows.

TABLE 26
A Hypothetical Factor Matrix

Test	Factor I	Factor II
1. Vocabulary	.74	.54
2. Analogies	.64	.39
3. Sentence Completion	.68	.43
4. Disarranged Sentences	.32	.23
5. Reading Comprehension	.70	.50
6. Addition	.22	–.51
7. Multiplication	.40	–.50
8. Arithmetic Problems	.52	–.48
9. Equation Relations	.43	–.37
10. Number Series Completion	.32	–.25

Several different methods for analyzing a set of variables into common factors have been derived. As early as 1901, Pearson (1901) pointed the way for this type of analysis; and Spearman (1904, 1927) developed a precursor of modern factor analysis. Kelley (1935) and Thurstone (1947) in America and Burt (1941) in England did much to advance the method. Alternative procedures, modifications, and refinements have been developed by many others. The availability of electronic computers is rapidly leading to the adoption of more refined and laborious techniques. Although differing in their initial postulates, most of these methods yield similar results.[1] For brief and simple introductions to the specific procedures of factor analysis, the reader is referred to Guilford (1954, Ch. 16) and Adcock (1954). A comprehensive and much more advanced treatment can be found in Harman (1967).

It is beyond the scope of this book to cover the mathematical basis or the computational procedures of factor analysis. An understanding of the results of factor analysis, however, need not be limited to those who have mastered its specialized methodology. Even without knowing how the factor loadings were computed, it is possible to see how a factor matrix is utilized in the identification and interpretation of factors. For an intelligent reading of reports of factor-analytic research, however, familiarity with a few other concepts and terms is helpful.

---

[1] The factor matrix shown in Table 26 is typical of those obtained with the centroid method, which was widely used prior to the advent of high-speed computers. Though crude, this method yields results that closely approximate those of the more refined and currently popular method of principal axes.

THE REFERENCE AXES. It is customary to represent factors geometrically as reference axes in terms of which each test can be plotted. Figure 58 illustrates this procedure. In this graph, each of the 10 tests from Table 26 has been plotted against the two factors, which correspond to axes I and II. Thus, the point representing Test 1 is located by moving .74 of the distance along axis I and .54 of the distance along axis II. The points corresponding to the remaining 9 tests are plotted in the same way, using the weights given in Table 26. Although all the weights on Factor I are positive, it will be noted that on Factor II some of the weights are positive and some negative. This can also be seen in Figure 58, where Tests 1 to 5 cluster in one part of the graph and Tests 6 to 10 in another.

In this connection it should be noted that the position of the reference axes is not fixed by the data. The original correlation table determines only the position of the tests (points in Figure 58) *in relation to each other*. The same points can be plotted with the reference axes in any position. For this reason, factor analysts usually rotate axes until they obtain the most satisfactory and easily interpretable pattern. This is a legitimate

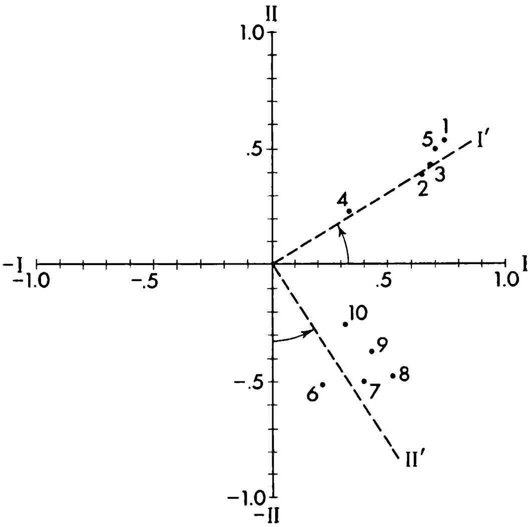


Fig. 58. A Hypothetical Factor Pattern, Showing Weights of Two Group Factors in Each of Ten Tests.

procedure, somewhat analogous to measuring longitude from, let us say, Chicago rather than Greenwich.

The reference axes in Figure 58 were rotated to positions I′ and II′, shown by the broken lines.[2] This rotation was carried out in accordance with Thurstone's criteria of *positive manifold* and *simple structure.* The former requires the rotation of axes to such a position as to eliminate all significant negative weights. Most psychologists regard negative factor loadings as inapplicable to aptitude tests, since such a loading implies that the higher the individual rates in the particular factor, the poorer will be his performance on the test. The criterion of simple structure means that each test shall have loadings on as few factors as possible. Both of these criteria are designed to yield factors that can be most readily and unambiguously interpreted. If a test has a high loading on a single factor and no significant loading on any other, we can learn something about the nature of the factor by examining the content of the test. If instead the test has moderate to low loadings on six factors, it can tell us little about the nature of any of these factors.

TABLE 27
Rotated Factor Matrix

(Data from Figure 58)

Test	Factor I	Factor II
1. Vocabulary	.91	−.06
2. Analogies	.75	.02
3. Sentence Completion	.80	.00
4. Disarranged Sentences	.39	−.02
5. Reading Comprehension	.86	−.04
6. Addition	−.09	.55
7. Multiplication	.07	.64
8. Arithmetic Problems	.18	.68
9. Equation Relations	.16	.54
10. Number Series Completion	.13	.38

It will be seen that on the rotated axes in Figure 58 all the verbal tests (Tests 1 to 5) fall along or very close to axis I′. Similarly, the numerical tests (Tests 6 to 10) cluster closely around axis II′. The new factor loadings, measured along the rotated axes, are given in Table 27. The reader may easily verify these factor loadings by preparing a paper "ruler" with a scale of units corresponding to that in Figure 58. With this ruler, distances can be measured along the rotated axes. The factor loadings in Table 27 include no negative values except for very low, negligible amounts attributable to sampling errors. All of the verbal tests have high loadings on Factor I′ and practically zero loadings on Factor II′. The numerical tests, on the other hand, have high loadings on Factor II′ and low, negligible loadings on Factor I′. The identification and naming of the two factors and the description of the factorial composition of each test have thus been simplified by the rotation of reference axes. In actual practice, the number of factors is often greater than two—a condition that complicates the geometrical representation as well as the statistical analysis, but does not alter the basic procedure. The fact that factor patterns are rarely as clear-cut as the one illustrated in Figure 58 adds further to the difficulty of rotating axes and of identifying factors.

---

[2] The reader may feel that rotated axis II′ should have been labeled −II′, to correspond to the unrotated axis −II. Which pole of the axis is labeled plus and which minus, however, is an arbitrary matter. In the present example, the rotated axis II′ has been "reflected" in order to eliminate negative weights.

Some factor analysts use a theoretical model as a guide in the rotation of axes. Invariance, or the corroboration of the same factors in independent but comparable investigations, is also taken into account. The rotated patterns resulting from such procedures usually exhibit simple structure (Guilford & Hoepfner, 1971).

INTERPRETATION OF FACTORS. Once the rotated factor matrix has been computed, we can proceed with the interpretation and naming of factors. This step calls for psychological insight rather than statistical training. To learn the nature of a particular factor, we simply examine the tests having high loadings on that factor and we try to discover what psychological processes they have in common. The more tests there are with high loadings on a given factor, the more clearly we can define the nature of the factor. In Table 27, for example, it is apparent that Factor I′ is verbal and Factor II′ is numerical.

The factor loadings given in Table 27 also represent the correlation of each test with the factor.[3] It will be recalled that this correlation is the factorial validity of the test (Ch. 6). From Table 27 we can say, for instance, that the factorial validity of the Vocabulary test as a measure of the verbal factor is .91. The factorial validity of the Addition test, in terms of the numerical factor, is .55. Obviously the first five tests have negligible validity as measures of the numerical factor, and the last five have practically no validity as measures of the verbal factor. The concept of factorial validity is especially relevant to the type of tests to be discussed in this chapter.

---

[3] This is true only when an orthogonal rotation of axes is applied. With oblique rotation, to be defined in a later section of this chapter, factor loadings and factor correlations bear a simple relation to each other and each can be found from the other by appropriate computations.

FACTORIAL COMPOSITION OF A TEST. One of the basic theorems of factor analysis states that the total variance of a test is the sum of the variances contributed by the common factors (shared with other tests) and the specific factors (occurring in that test alone), plus the error variance. We have already encountered error variance in the analysis of test scores (Ch. 5). If, for instance, the reliability coefficient of a test is .83, we conclude that 17 percent of the variance of scores on this test is error variance (1.00 − .83 = .17). Through factor analysis, we can further subdivide the sources of variance contributing toward performance on any test.

TABLE 28
Source of Variance of Test Scores

Test	Common Factor Loadings			Reliability Coefficient	Proportional Contribution				
	V	N	R		V	N	R	Specific	Error
1. Arithmetic Reasoning	.40	.55	.60	.90	.16	.30	.36	.08	.10
2. Multiplication	.10	.70	.30	.85	.01	.49	.09	.26	.15

Let us consider the two tests listed in Table 28. For each test, we have its factor loading in Verbal (V), Numerical (N), and Reasoning (R) factors, as well as its reliability coefficient. Since each factor loading also represents the correlation between the test and the factor, the square of the factor loading gives us the proportion of common variance between the test and that factor. In the last section of the table, each factor loading has been squared to show the proportional contribution of that factor to the total variance of test scores. Thus, we find that in the Arithmetic Reasoning test 16 percent of the variance is attributable to the Verbal factor, 30 percent to the Numerical factor, and 36 percent to the Reasoning factor. The error variance in the last column is found by simply subtracting the reliability coefficient from the total variance (1.00 − .90 = .10). Whatever is left represents the specificity of this test, i.e., that portion of its "true" variance it does not share with any other test with which it was factor analyzed. For the Arithmetic Reasoning test, we have:

$$.16 + .30 + .36 + .10 = .92$$
$$1.00 - .92 = .08$$

Figure 59 provides a pictorial breakdown of the sources of variance for the two tests of Table 28.

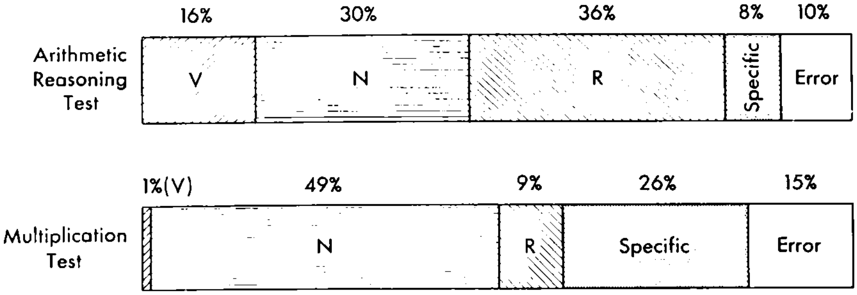


FIG. 59. Percentage of Common, Specific, and Error Variance in Two Hypothetical Tests.

(Data from Table 28.)

Any individual's performance on these two tests depends on the amounts of each of the relevant abilities or factors he possesses, as well as the relative weights of these factors in the particular test. Thus, if we had the individual's score in the V, N, and R factors, expressed in the same units, we could weight each score by multiplying it by the corresponding factor loading. The sum of these products would provide an estimate of the individual's score on the test. The smaller the contribution of specific and error factors to the test, the more accurate would this estimate be.

In the example given in Table 28, if an individual rates very high in V, this would help him much more on the Arithmetic Reasoning than on the Multiplication test. In fact, it would help him four times as much, since the weight of the V factor is four times as great in Arithmetic Reasoning as in Multiplication (.40 *vs.* .10). Of the three common factors, N would have the greatest effect on Multiplication (loading = .70) and R would have the greatest effect on Arithmetic Reasoning (loading = .60).

FACTOR LOADINGS AND CORRELATION. A second basic theorem of factor analysis concerns the relationship between factor loadings and the correlations among variables. The correlation between any two variables is equal to the sum of the cross-products of their common-factor loadings. Since specific and error factors are unique to each variable, they cannot contribute to the correlation between variables. The correlation between any two variables depends only on the factors that these two variables share. The larger the weights of these common factors in the two variables, the higher will be the correlation between the variables. The correlation between the two tests given in Table 28 can be found by multiplying the loadings of each of the three common factors in the two tests and adding the products, as shown below:

$$r_{12} = (.40)(.10) + (.55)(.70) + (.60)(.30) = .60$$

OBLIQUE AXES AND SECOND-ORDER FACTORS. The axes employed in Figure 58 are known as *orthogonal axes,* since they are at right angles to each other. Occasionally, the test clusters are so situated that a better fit can be obtained with *oblique axes.* In such a case, the factors would themselves be correlated. Some investigators have maintained that orthogonal, or uncorrelated, factors should always be employed, since they provide a simpler and clearer picture of trait relationships. Others insist that oblique axes should be used when they fit the data better, since the most meaningful categories need not be uncorrelated. An example cited by Thurstone is that of height and weight. Although it is well known that height and weight are highly correlated, they have proved to be useful categories in the measurement of physique.

When the factors are themselves correlated, it is possible to subject the intercorrelations among the factors to the same statistical analyses we employ with intercorrelations among tests. In other words, we can "factorize the factors" and derive *second-order factors.* This process has been followed in a number of studies with both aptitude and personality variables. Certain investigations with aptitude tests have yielded a single second-order general factor. As a rule, American factor analysts have proceeded by accounting for as much of the common variance as possible through group factors and then identifying a general factor as a second-order factor if the data justified it. British psychologists, on the other hand, usually begin with a general factor, to which they attribute the major portion of the common variance, and then resort to group factors to account for any remaining correlation. These procedural differences reflect differences in theoretical orientation to be discussed in the following section.

THEORIES OF TRAIT ORGANIZATION

THE TWO-FACTOR THEORY. The first theory of trait organization based on a statistical analysis of test scores was the two-factor theory developed by the British psychologist Charles Spearman (1904, 1927). In its original formulation, this theory maintained that all intellectual activities share a single common factor, called the *general factor,* or *g.* In addition, the theory postulated numerous *specifics,* or *s* factors, each being strictly specific to a single activity. Positive correlation between any two functions was thus attributed to the *g* factor. The more highly the two functions were "saturated" with *g*, the higher would be the correlation between them. The presence of specifics, on the other hand, tended to lower the correlation between functions.

Although two types of factors, general and specific, are posited by this theory, it is only the single factor *g* that accounts for correlation. In contrast to other theories of trait relations, therefore, it could be more precisely characterized as a single-factor theory, although the original designation has persisted. Figure 60 illustrates the basis for correlation among tests according to this theory. In this illustration, tests 1 and 2 would correlate highly with each other since each is highly saturated with *g*, as shown by the shaded areas. The white area in each test represents specific and error variance. Test 3 would have low correlations with each of the other two tests, since it contains very little *g*.

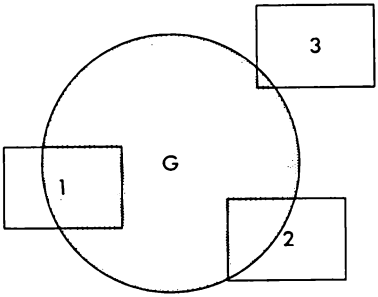


Fig. 60. Correlation Model Underlying Two-Factor Theory.

It follows from the two-factor theory that the aim of psychological testing should be to measure the amount of each individual's *g*. If this factor runs through all abilities, it furnishes the only basis for prediction of the individual's performance from one situation to another. It would be futile to measure specific factors, since each by definition operates in only a single activity. Accordingly, Spearman proposed that a single test, highly saturated with *g*, be substituted for the heterogeneous collection of items found in intelligence tests. He suggested that tests dealing with abstract relations are probably the best measures of *g* and could be used for this purpose. Examples of tests constructed as measures of *g* include Raven's Progressive Matrices and Cattell's Culture Fair Intelligence Test, both discussed in Chapter 10.

From the outset, Spearman realized that the two-factor theory must be qualified. When the activities compared are very similar, a certain degree of correlation may result over and above that attributable to the *g* factor.

Thus, in addition to general and specific factors, there might be another, intermediate class of factors, not so universal as *g* nor so strictly specific as the *s* factors. Such a factor, common to a group of activities but not to all, has been designated as a *group factor*. In the early formulation of his theory Spearman admitted the possibility of very narrow and negligibly small group factors. Following later investigations by several of his students, he included much broader group factors such as arithmetic, mechanical, and linguistic abilities.

MULTIPLE-FACTOR THEORIES. The prevalent contemporary American view of trait organization recognizes a number of moderately broad group factors, each of which may enter with different weights into different tests. For example, a verbal factor may have a large weight in a vocabulary test, a smaller weight in a verbal analogies test, and a still smaller weight in an arithmetic reasoning test. Figure 61 illustrates the intercorrelations among five tests in terms of a multiple-factor model. The correlations of tests 1, 2, and 3 with each other result from their common loadings with the verbal factor (V). Similarly, the correlation between tests 3 and 5 results from the Spatial factor (S), and that between tests 4 and 5 from the Number factor (N). Tests 3 and 5 are factorially complex, each having appreciable loadings in more than one factor: V and S in test 3, N and S in test 5. From the second basic theorem of factor analysis, discussed in the preceding section, we can also tell something about the relative magnitude of the intercorrelations. For example, test 3 will correlate higher with test 5 than with test 2 because the weights of the S factor in tests 3 and 5 (represented by the overlapping areas) are larger than the weights of the V factor in tests 2 and 3.

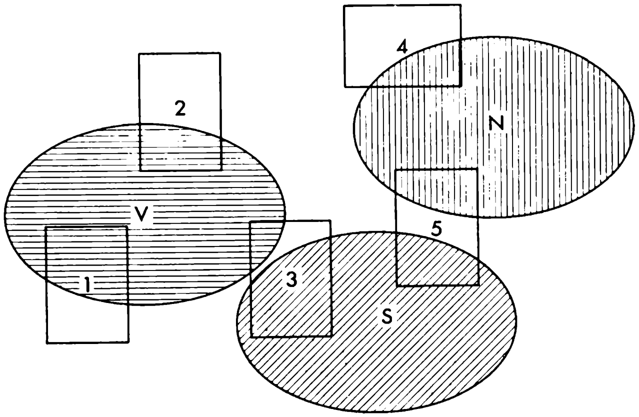


FIG. 61. Correlation Model Underlying Multiple-Factor Theories.

The publication of Kelley's *Crossroads in the Mind of Man* (1928) paved the way for a large number of studies in quest of particular group factors. Chief among the factors proposed by Kelley were manipulation of spatial relationships, facility with numbers, facility with verbal material, memory, and speed. This list has been modified and extended by subsequent investigators employing the more modern methods of factor analysis discussed in the preceding section.

One of the leading exponents of multiple-factor theory was Thurstone. On the basis of extensive research by himself and his students, Thurstone proposed about a dozen group factors which he designated as "primary mental abilities." Those most frequently corroborated in the work of Thurstone and of other independent investigators (French, 1951; Thurstone, 1938; Thurstone & Thurstone, 1941) include the following:

V. *Verbal Comprehension:* The principal factor in such tests as reading comprehension, verbal analogies, disarranged sentences, verbal reasoning, and proverb matching. It is most adequately measured by vocabulary tests.

W. *Word Fluency:* Found in such tests as anagrams, rhyming, or naming words in a given category (e.g., boys' names, words beginning with the letter T).

N. *Number:* Most closely identified with speed and accuracy of simple arithmetic computation.

S. *Space:* This factor may represent two distinct factors, one covering perception of fixed spatial or geometric relations, the other manipulatory visualizations, in which changed positions or transformations must be visualized.

M. *Associative Memory:* Found principally in tests demanding rote memory for paired associates. There is some evidence to suggest that this factor may reflect the extent to which memory crutches are utilized (Christal, 1958). The evidence is against the presence of a broader factor through all memory tests. Other restricted memory factors, such as memory for temporal sequences and for spatial position, have been suggested by some investigations.

P. *Perceptual Speed:* Quick and accurate grasping of visual details, similarities, and differences. This factor may be the same as the speed factor identified by Kelley and other early investigators. This is one of several factors subsequently identified in perceptual tasks (Thurstone, 1944).

I (or *R*). *Induction* (or *General Reasoning*): The identification of this factor was least clear. Thurstone originally proposed an inductive and a deductive factor. The latter was best measured by tests of syllogistic reasoning and the former by tests requiring the subject to find a rule, as in a number series completion test. Evidence for the deductive

> factor, however, was much weaker than for the inductive. Moreover, other investigators suggested a general reasoning factor, best measured by arithmetic reasoning tests.

It should be noted that the distinction between general, group, and specific factors is not so basic as may at first appear. If the number or variety of tests in a battery is small, a single general factor may account for all the correlations among them. But when the same tests are included in a larger battery with a more heterogeneous collection of tests, the original general factor may emerge as a group factor, common to some but not all tests. Similarly, a certain factor may be represented by only one test in the original battery, but may be shared by several tests in the larger battery. Such a factor would have been identified as a specific in the original battery, but would become a group factor in the more comprehensive battery. In the light of these considerations, it is not surprising to find that intensive factorial investigations of special areas have yielded many factors in place of the one or two primary mental abilities originally identified in each area. Such has been the case in studies of verbal, perceptual, memory, and reasoning tests.

Factorial research seems to have produced a bewildering multiplication of factors. The number of cognitive factors reported to date by different investigators is well over 100. A certain amount of order has been achieved by cross-identification of factors reported by different investigators and often given different names (French, 1951). Such cross-identification can be accomplished when there are several tests common to the investigations being compared. To facilitate this process, a group of factor analysts assembled a kit of "reference tests" measuring the principal aptitude factors so far identified. This kit, which is distributed by Educational Testing Service (French, Ekstrom, & Price, 1963), makes it easier for different investigators planning factorial research to include some common tests in their batteries. The kit covers 24 cognitive factors, with two to five tests for measuring each factor.

It is apparent that even after these efforts at simplification and coordination, the number of factors remains large. Human behavior is varied and complex, and perhaps it is unrealistic to expect a dozen or so factors to provide an adequate description of it. For specific purposes, however, we can choose appropriate factors with regard to both nature and breadth. For example, if we are selecting applicants for a difficult and highly specialized mechanical job, we would probably want to measure fairly narrow perceptual and spatial factors that closely match the job requirements. In selecting college students, on the other hand, a few broad factors such as verbal comprehension, numerical facility, and general reasoning would be most relevant. Illustrations of the different ways in which factorial results have been utilized in test development will be given later in this chapter.

STRUCTURE-OF-INTELLECT MODEL. Some factor analysts have tried to simplify the picture of trait relationships by organizing the traits into a systematic schema. On the basis of more than two decades of factor-analytic research, Guilford (1967; Guilford & Hoepfner, 1971) has proposed a boxlike model which he calls the structure-of-intellect model (SI). Illustrated in Figure 62, this model classifies intellectual traits along three dimensions:

Operations—what the respondent does. These include cognition, memory, divergent production (prominent in creative activity), convergent production, and evaluation.

Contents—the nature of the materials or information on which operations are performed. These include figural, symbolic (e.g., letters, numbers), semantic (e.g., words), and behavioral (information about other persons' behavior, attitudes, needs, etc.).

Products—the form in which information is processed by the respondent. Products are classified into units, classes, relations, systems, transformations, and implications.

Since this classification includes 5 × 4 × 6 categories, there are 120 cells in the model. In each cell, at least one factor or ability is expected; some

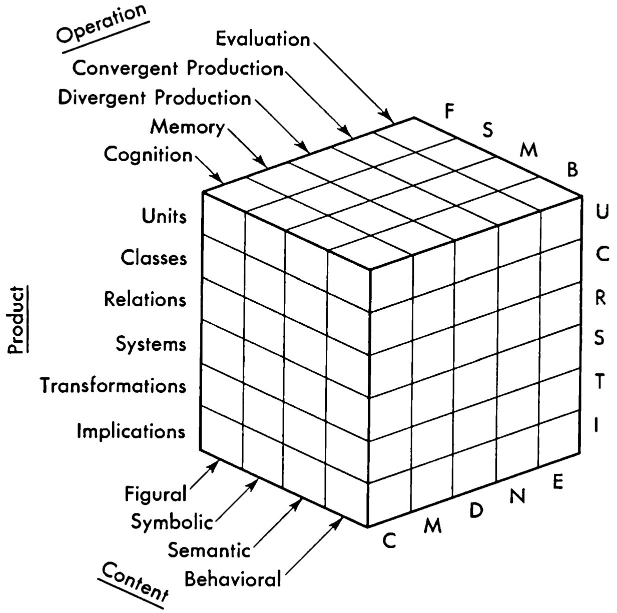

Operation
Evaluation
Convergent Production
Divergent Production
Memory
Cognition
F
S
M
B
Units
Classes
Relations
Systems
Transformations
Implications
Product
U
C
R
S
T
I
Figural
Symbolic
Semantic
Behavioral
Content
C
M
D
N
E

FIG. 62. Three-Dimensional Model of the Structure of Intellect.
(From Guilford, 1967, p. 63.)

cells may contain more than one factor. Each factor is described in terms of all three dimensions. Figure 62 gives both the names and letter symbols for each dimension. The code used by Guilford and his associates in describing each factor specifies Operation, Content, and Product in that order. The well-known factor of verbal comprehension, for example, corresponds to cognition of semantic units (CMU) and is best measured by vocabulary tests. A memory span test utilizing series of unrelated letters or numbers assesses memory for symbolic units (MSU). Upon the completion of the Aptitudes Research Project—a 20-year coordinated program of research on the SI model—Guilford and his associates had identified 98 of the anticipated factors (Guilford & Hoepfner, 1971). The remaining unfilled cells provide a plan for further research. An alphabetical list and brief description of the many tests developed in the course of this 20-year project can be found in the book by Guilford and Hoepfner (1971, Appendix B).[4] It may also be noted that Meeker (1969) has applied the SI categories in classifying the items of the Stanford-Binet and Wechsler scales.

HIERARCHICAL THEORIES. An alternative schema for the organization of factors has been proposed by a number of British psychologists, including Burt (1949) and Vernon (1960), and by Humphreys (1962) in America. A diagram illustrating Vernon's application of this system is reproduced in Figure 63. At the top of the hierarchy, Vernon places Spearman's g factor. At the next level are two broad group factors, corresponding to verbal-educational (*v*:*ed*) and to practical-mechanical (*k*:*m*) aptitudes,

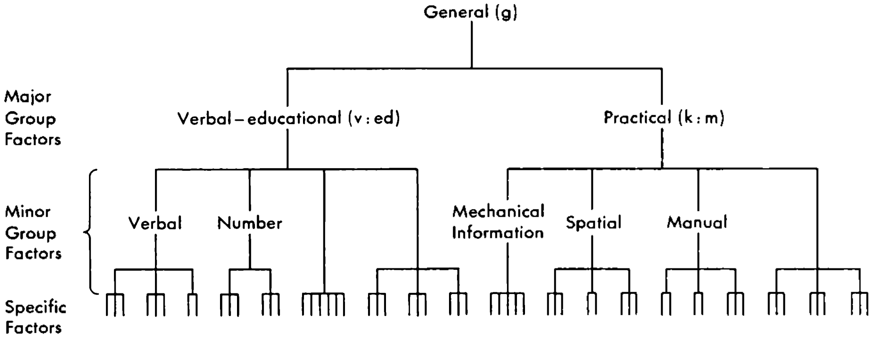


Fig. 63. Model of a Hierarchical Organization of Abilities.

(Adapted from Vernon, 1960, p. 22. Copyright 1960, Methuen & Co., Ltd.)

---

[4] About 30 of these tests have thus far been published and can be purchased by qualified persons from Sheridan Psychological Services.

respectively. These major factors may be further subdivided. The verbal-educational factor, for example, yields verbal and numerical subfactors, among others. Similarly, the practical-mechanical factor splits into such subfactors as mechanical-information, spatial, and manual. Still narrower subfactors can be identified by further analysis, let us say, of the verbal tasks. At the lowest level of the hierarchy are the specific factors. Such a hierarchical structure thus resembles an inverted genealogical tree, with *g* at the top, *s* factors at the bottom, and progressively narrower group factors in between.

Humphreys (1962, 1970) also recommends a hierarchical model as a means of coping with the proliferation of factors. Rather than considering the first-order, narrower factors as primary, he suggests that each test constructor or user choose that level of the hierarchy that is most appropriate for his purposes. Humphreys recognizes, however, that a single test may be classified into more than one hierarchy, with reference to content, method, and other facets. To measure any one facet, he proposes that the test be made *heterogeneous* with regard to all other facets. For example, if we are interested in the subject's ability to solve analogies problems, we should use a test that includes verbal, numerical, pictorial, and spatial analogies. If we wish to measure verbal ability, we should do so with a variety of item types, such as vocabulary, analogies, and series completion. This procedure contrasts with that followed by Guilford, who seeks separate factors (and tests) for each homogeneous cell in his three-way classification.

NATURE AND DEVELOPMENT OF FACTORS. That different investigators may arrive at dissimilar models of trait organization becomes less perplexing when we recognize that the traits identified through factor analysis are simply an expression of correlation among behavior measures. They are not underlying entities or causal factors, but descriptive categories. Hence, it is conceivable that different principles of classification may be applicable to the same data.

The concept of factors as descriptive categories is explicit in the writings of Thomson (1916, 1948), Burt (1941, 1944), and Vernon (1960) in England, and those of Tryon (1935) in America. All these writers called attention to the vast multiplicity of behavioral elements, which may become organized into clusters through either hereditary or experiential linkages. For example, persons in our culture are likely to develop a broad verbal-educational factor running through all activities learned in school. A narrower factor of numerical aptitude may result from the fact that all arithmetic processes are taught together by the same teacher in the same classroom. Hence, the child who is discouraged, antagonized, or bored during the arithmetic period will tend to fall behind in his learning of *all* these processes; the one who is stimulated and gratified in his arithmetic class will tend to learn well all that is taught in that class period and to develop attitudes that will advance his subsequent numerical learning.

There has been an increasing recognition of the role of experiential conditions in the formation of group factors. It is not only the level of performance in different abilities, but also the way in which performance is organized into distinct traits that is influenced by experiential background. Differences in factor patterns have been found to be associated with different cultures or subcultures, socioeconomic levels, and types of school curricula (see Anastasi, 1970). Changes in factor patterns over time are also relevant. These include long-term changes—which may reflect the cumulative effect of everyday-life experiences—as well as short-term changes resulting from practice and other experimentally controlled learning experiences (Fleishman, 1972). Research on animals has also yielded suggestive evidence regarding the experimental production of factors by the control of early experiences (Whimbey & Denenberg, 1966).

The factorial composition of the same objective task may differ among individuals with diverse experiential backgrounds. One reason for these individual differences may be found in the use of different methods to carry out the same task. Individuals with highly developed verbal abilities, for example, will tend to utilize verbal mediators to solve a mechanical or spatial problem; those whose experiences have been predominantly mechanical, on the other hand, will tend to follow a perceptual or spatial approach in solving the same problem. Relevant evidence is provided by French (1965), who found that the factorial composition of the same tests differed between groups of persons classified according to their typical problem-solving styles. Suggestive evidence can also be found in a study by C. H. Frederiksen (1969) on the cognitive strategies subjects use in memorizing words. In the course of learning, an individual may change his choice of strategy and thereby alter the factorial composition of the task as performed by him.

A mechanism for the emergence of factors is provided by the familiar concepts of learning set and transfer of training (J. B. Carroll, 1966; Ferguson, 1954, 1956; Whiteman, 1964). The establishment of learning sets enables the individual to learn more efficiently when presented with a new problem of the same kind. In Harlow's (1949, 1960) classic experiments with monkeys, after the animal had solved problems requiring the differentiation between certain shapes (like triangle and circle), it learned to discriminate between *other* shapes much more rapidly than did animals without such prior experience. The animal had established a learning set for differentiating shapes; he knew what to look for when faced with a new problem. He had thus "learned how to learn" this type of problem.

Similarly, many of the skills developed through formal schooling, such as reading and arithmetic computation, are applicable to a wide variety of subsequent learning situations. Efficient and systematic problem-solving techniques can likewise be applied to the solution of new problems. Individual differences in the extent to which these skills have been acquired will thus be reflected in the performance of a large number of different tasks; and in a factor analysis of these tasks, these widely applicable skills would emerge as broad group factors. The breadth of the transfer effect, or the variety of tasks to which a skill is applicable, would thus determine the breadth of the resulting group factor.

In summary, the factors or abilities identified through factor analysis are descriptive categories, reflecting the changing interrelationships of performance in a variety of situations. These factors are not static entities but are themselves the product of the individual's cumulative experiential history. Insofar as the interrelationships of experiences vary among individuals and groups, different factor patterns may be expected among them. As the individual's experiences change—through formal education, occupational functions, and other continuing activities—new traits may become differentiated or previously existing traits may merge into broader composites.

MULTIPLE APTITUDE BATTERIES

BATTERIES FOR GENERAL USE. One of the major effects of factor analysis on test construction can be seen in the development of multiple aptitude batteries. These batteries yield a profile of test scores in a set of relatively independent abilities identified through factor-analytic research. The Chicago Tests of Primary Mental Abilities (PMA) represent the first systematic effort to construct such a battery. Originally published in 1941, this battery was the direct outcome of Thurstone's previously described factor-analytic investigations. For each factor, Thurstone selected those tests with the highest factorial validities. This battery, designed principally for the high school and college levels, required six testing sessions.

Later versions of the PMA tests, published by Science Research Associates, were characterized chiefly by reduction in length and extension to younger age levels. These tests were marred by a number of technical deficiencies, which have been fully discussed in a series of reviews in the *Mental Measurements Yearbooks,* as well as in other critical surveys (e.g., Super, 1958). The chief criticisms pertained to inadequacies of normative data, questionable types of scores (such as ratio IQ's), unsupported interpretations of scores in terms of educational and vocational criteria, meager validity data, improper procedures for computing the reliability of speeded tests, excessive dependence of scores on speed, and low reliabilities of certain factor scores.

While several of these deficiencies were corrected in the 1962 edition, the PMA tests still fall short of the technical standards set by other contemporary instruments of the same type. The original Thurstone tests were based on extensive factor-analytic research and represented an important breakthrough in test construction. The expectations aroused by the initial experimental test series, however, were not realized in the commercially published batteries. As one reviewer expressed it:

> While the history of PMA as a test has been one of steady decline, the history of primary mental abilities as a paradigm has been one of steady growth. The undergirding conceptual framework of PMA provided a common paradigm which not only generated competing multifactor instruments but also structured the "normal science" within which such tests are still being viewed and used (Schutz, 1972).

One of the most widely used multiple aptitude batteries is the Differential Aptitude Tests (DAT). First published in 1947, the DAT has been revised in 1963 and in 1973. This battery was designed principally for use in the educational and vocational counseling of students in grades 8 to 12. Although not utilizing factor analysis in its construction, the authors of the DAT were guided in their choice of tests by the accumulated results of factorial research, as well as by practical counseling needs. Rather than striving for factorial purity, they included tests that might be factorially complex if they represented well-recognized vocational or educational areas. The DAT yields the following eight scores: Verbal Reasoning, Numerical Ability, Abstract Reasoning, Clerical Speed and Accuracy, Mechanical Reasoning, Space Relations, Spelling, and Language Usage. A sample item from each test is reproduced in Figures 64A and 64B. The complete battery is available in two equivalent forms, S and T.

The DAT manual provides a full account of the test-construction procedures followed in developing the battery. Norms were derived from over 64,000 students in 76 public and parochial school systems distributed over 33 states and the District of Columbia. The standardization sample was selected by a two-stage sampling procedure so as to be representative of the entire student population of the United States in grades 8 to 12 with respect to socioeconomic level of the community, size of enrollment of the school district, and scholastic achievement and ethnic composition of the students.

With the exception of Clerical Speed and Accuracy, all DAT tests are

VERBAL REASONING

Choose the correct pair of words to fill the blanks. The first word of the pair goes in the blank space at the beginning of the sentence; the second word of the pair goes in the blank at the end of the sentence.

is to night as breakfast is to

A. supper —— corner
B. gentle —— morning
C. door —— corner
D. flow —— enjoy
E. supper —— morning

The correct answer is E.

NUMERICAL ABILITY

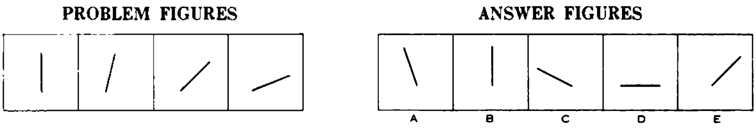
Choose the correct answer for each problem.

Add	**13**	**A 14**	**Subtract**	**30**	**A 15**
	12	**B 25**		**20**	**B 26**
		C 16			**C 16**
		D 59			**D 8**
		N none of these			**N none of these**

The correct answer for the first problem is B; for the second, N.

ABSTRACT REASONING

The four "problem figures" in each row make a series. Find the one among the "answer figures" that would be next in the series.



The correct answer is D

CLERICAL SPEED AND ACCURACY

In each test item, one of the five combinations is underlined. Find the same combination on the answer sheet and mark it.

TEST ITEMS

V.	AB	AC	AD	AE	AF
W.	aA	aB	BA	Ba	Bb
X.	A7	7A	B7	7B	AB
Y.	Aa	Ba	bA	BA	bB
Z.	3A	3B	33	B3	BB

SAMPLE OF ANSWER SHEET

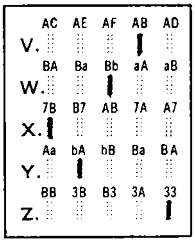
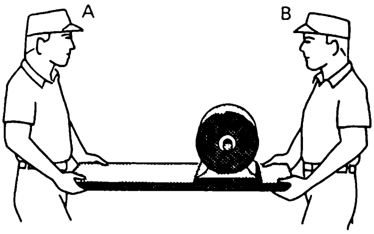


Fig. 64A. Sample Items from the Differential Aptitude Tests.

(Reproduced by permission. Copyright © 1972 by The Psychological Corporation, New York, N.Y. All rights reserved.)

MECHANICAL REASONING

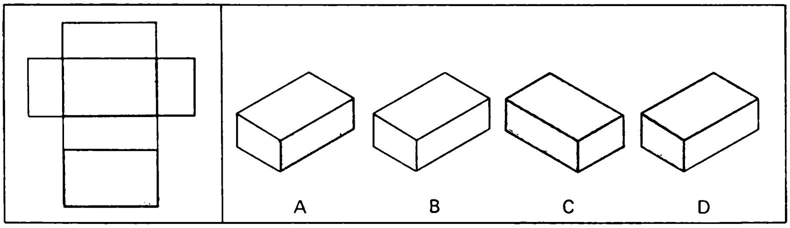
Which man has the heavier load? (If equal, mark C.)



The correct answer is B.

SPACE RELATIONS

Which one of the following figures could be made by folding the pattern at the left? The pattern always shows the outside of the figure. Note the grey surfaces.



The correct answer is D.

SPELLING

Indicate whether each word is spelled right or wrong.

W. man

X. gurl

	R	W
W.	▮	
	R	W
X.	⁞	▮

LANGUAGE USAGE

Decide which of the lettered parts of the sentence contains an error and mark the corresponding letter on the answer sheet. If there is no error, mark N.

X. Ain't we / going to / the office / next week?
A B C D

	A	B	C	D	N
X.	▮	⁞	⁞	⁞	⁞

Fig. 64B. Sample Items from the Differential Aptitude Tests.

(Reproduced by permission. Copyright © 1972 by The Psychological Corporation, New York, N.Y. All rights reserved.)

essentially power tests. Reliability coefficients are high and permit interpretation of intertest differences with considerable confidence. By combining information on test reliabilities with the intercorrelations among tests, the test authors determined the proportion of differences in excess of chance between each pair of tests. These percentages ranged from 25 to 57, most falling in the 30's and 40's. In other words, for all possible test pairs, from one fourth to one half of the examinees in representative samples obtained larger score differences between tests than would be expected from errors of measurement.

Both percentile and stanine norms are provided. Individual percentile scores are plotted on normalized percentile charts, thereby eliminating the inequalities of percentile units (see Fig. 5, Ch. 4). In order to facilitate the interpretation of score differences between tests, the profile charts show a percentile band for each score, such that the chances are approximately 90 out of 100 that the individual's "true" score lies within the band (see Fig. 13, Ch. 5). As was explained in Chapter 5, test scores whose percentile bands do not overlap can be regarded as significantly different.

The amount of validity data available on the DAT is overwhelming, including several thousand validity coefficients. Most of these data are concerned with predictive validity in terms of high school achievement in both academic and vocational programs. Many of the coefficients are high, even with intervals as long as three years between test and criterion data. The results are somewhat less encouraging with regard to differential prediction. Although, in general, verbal tests correlate more highly with English courses and numerical tests with mathematics courses, there is evidence of a large general factor underlying performance in all academic work. Verbal Reasoning, for example, gives high correlations with most courses. It is chiefly for this reason that the VR + NA score was introduced as an index of scholastic aptitude. Being the sum of the raw scores on the Verbal Reasoning and Numerical Ability subtests, this index correlates in the .70's and .80's with composite criteria of academic achievement. Norms are provided for this index, which is one of the scores regularly included in the DAT profile (see Fig. 13, Ch. 5). The manual also provides regression equations for predicting the Verbal and Mathematical scores on the CEEB Scholastic Aptitude Test from appropriate combinations of three DAT subtests.

With regard to vocational criteria, there is some evidence for the predictive validity of individual DAT subtests, but the data are relatively meager. A four-year follow-up of 1,700 high school juniors and seniors and a seven-year follow-up of a somewhat smaller group revealed several profile differences among students who subsequently entered different educational curricula or occupations.

As an aid to educational and vocational counseling, there is also a DAT Career Planning Program which utilizes DAT scores and student responses to a Career Planning Questionnaire. In the questionnaire, the student records his liking for various school subjects and other school activities, his educational and vocational goals, his general academic performance, and his interest in different fields of work and representative occupations. The DAT Career Planning Report is a computer printout including the student's DAT scores, with appropriate explanations, and interpretive statements that relate the DAT scores to the student's expressed interests and plans. The computer program was prepared by vocational psychologists on the basis of the available fund of information about job requirements and about the occupational validity of aptitude tests and interests. The manual for the DAT Career Planning Program (Super, 1973) gives a concise but clear description of the development of the program.

About a dozen multiple aptitude batteries have been developed for use in educational testing, counseling, and personnel classification. These instruments vary widely in approach, technical quality, and amount of available validation data. A common feature, however, is their disappointing performance with regard to differential validity. Yet it is in terms of differential validity that the distinctive contribution of multiple factor batteries should be evaluated. It is in connection with classification decisions that these batteries find their chief practical applications. Many have been designed explicitly for counseling purposes, in which classification decisions are preeminent. In a counseling situation, the profile of test scores is utilized to aid the counselee in choosing among several possible fields of educational or vocational specialization. Such a choice represents a typical classification decision. Institutional decisions regarding the assignment of personnel to different jobs or the admission of students to different educational curricula likewise involve a classification strategy. It will be recalled from our earlier discussion of validity (Ch. 7) that for classification instruments it is differential validity that is of prime importance.

Perhaps academic criteria, against which most multiple aptitude batteries have so far been validated, are not clearly differentiable on the basis of aptitudes. It is possible that differences in performance in specific courses depend principally on interests, motivation, and emotional factors. Unpredictable contingency factors, such as interpersonal relations between an individual student and a particular instructor, may also play a part. With regard to aptitudes, the large common contribution of verbal comprehension to achievement in all academic areas has been repeatedly demonstrated. If dependable criteria of occupational success (rather than performance in vocational *training*) can be developed, it is likely that better differential predictions can be made with batteries designed for this purpose. In terms of available data, however, multifactor batteries have fallen short of their initial promise.

BATTERIES FOR SPECIAL PROGRAMS. Factor analysis underlies the development of classification batteries widely employed in the armed services and in certain civilian agencies. The General Aptitude Test Battery (GATB) was developed by the United States Employment Service (USES) for use by employment counselors in the state employment service offices (Dvorak, 1956; U.S. Department of Labor, 1970a). Prior to the preparation of this battery, factorial analyses were conducted on preliminary batteries of 15 to 29 tests, which had been administered to nine groups of men. These groups included a total of 2,156 men between the ages of 17 and 39, most of whom were trainees in vocational courses. A total of 59 tests was covered by the nine overlapping batteries. On the basis of these investigations, 10 factors were identified and 15 tests chosen to measure them. In a later revision of the GATB, the number of tests was reduced to 12 and the number of factors to 9.

The factors currently covered by the GATB are as follows:

G. *General Learning Ability:* Found by adding the scores on three tests also used to measure other factors (Vocabulary, Arithmetic Reasoning, Three-Dimensional Space).

V. *Verbal Aptitude:* Measured by a Vocabulary test requiring examinee to indicate which two words in each set have either the same or opposite meaning.

N. *Numerical Aptitude:* Includes both Computation and Arithmetic Reasoning tests.

S. *Spatial Aptitude:* Measured by Three-Dimensional Space test, involving the ability to comprehend two-dimensional representation of three-dimensional objects and to visualize effects of movement in three dimensions.

P. *Form Perception:* Measured by two tests requiring the examinee to match identical drawings of tools in one test and of geometric forms in the other.

Q. *Clerical Perception:* Similar to *P*, but requiring the matching of names rather than pictures or forms.

K. *Motor Coordination:* Measured by a simple paper-and-pencil test requiring the examinee to make specified pencil marks in a series of squares.

F. *Finger Dexterity:* Two tests requiring the assembling and disassembling, respectively, of rivets and washers.

M. *Manual Dexterity:* Two tests requiring the examinee to transfer and reverse pegs in a board.

The four tests used for measuring factors *F* and *M* require simple apparatus; the other eight are paper-and-pencil tests. Alternate forms are available for the first seven tests, used for measuring factors *G* through *Q*. The entire battery requires approximately 2½ hours.

The nine factor scores on the GATB are converted into standard scores with a mean of 100 and an *SD* of 20. These standard score norms were derived from a sample of 4,000 cases representative of the 1940 working population of the United States in terms of age, sex, educational, occupational, and geographical distribution. This sample may thus be regarded as a fixed reference group to which the score scale is anchored. By testing many groups of employees, applicants, and trainees in different kinds of jobs, score patterns were subsequently established, showing the critical aptitudes and minimum standard scores required for each occupation. For example, accounting was found to require a minimum score of 105 in General Learning Ability (*G*) and 115 in Numerical Aptitude (*N*). Plumbing called for a minimum score of 85 in General Learning Ability (*G*), 80 in Spatial Aptitude (*S*) and in Manual Dexterity (*M*), and 75 in Clerical Perception (*Q*) (U.S. Department of Labor, 1970a, Section IV). An individual's standard score profile can thus be matched with all those occupations whose cutoff scores he reaches or exceeds, and these occupations can be considered in counseling him.

In USES terminology, the aptitudes (with their appropriate cutoff scores) chosen for a specific occupation constitute a Special Aptitude Test Battery (SATB). The development of an SATB was illustrated in Chapter 7. The procedure followed with each group includes job analysis, selection of suitable criterion data (output records, supervisors' ratings, training performance, etc.), and administration of the 12-test battery. The significant aptitudes are chosen on the basis of their criterion correlations, as well as the means and *SD*'s of scores on each aptitude and the qualitative job-analysis information. For example, if workers on the job under consideration average considerably above the normative sample in a particular aptitude and also show relatively low variability in their scores, that aptitude would probably be included in the SATB even if it fails to show a significant criterion correlation. Such a situation could occur if employees on a certain job were a highly selected group with regard to that aptitude.

To facilitate the use of score patterns in counseling, specific occupations having similar aptitude requirements have been grouped into a relatively small number of job families. Cutoff scores were then established in the three most significant aptitudes for each family. The resulting score patterns are designated as OAP's (Occupational Ability Patterns). By 1970, 62 OAP's had been prepared, covering several hundred specific occupations (U.S. Department of Labor, 1970a, Section II). More will be added as the need arises and additional data become available.

The GATB is used regularly in state employment service offices in the counseling and job referral of a vast number of persons. In addition, the battery may be obtained by nonprofit organizations such as secondary schools, colleges, Veterans' Administration hospitals, and prisons, under arrangements with the appropriate state employment service. Permission has also been granted to individuals and organizations in many foreign countries to translate the GATB and use it for research purposes.

Through the facilities of the USES and the various state offices, an impressive body of data has been accumulated on the GATB, and a continuing research program is in progress. Despite the brevity of individual tests, reliabilities of the factor scores are generally satisfactory. Both equivalent-form and retest correlations cluster in the .80's and low .90's, although the reliabilities of the motor tests tend to be somewhat lower.

A large amount of information on the validity of both SATB's and OAP's is reported in the GATB *Manual* (U.S. Department of Labor, 1970a, Section III; see also Bemis, 1968). Data were gathered through some 450 studies involving over 25,000 employees, applicants, apprentices, trainees, and students. Current validation studies include minority group representation and compute separate validities for minority and nonminority subsamples. Both predictive and concurrent validities have been investigated; but the large majority of the studies used only concurrent procedures. Most of the composite validities for the SATB's cluster around .40. Although cross validation has not been systematically carried out and rather crude statistical procedures have been employed, available cross validation data are encouraging. In an effort to evaluate the effectiveness of the GATB for counseling young people, Droege (1968) correlated the GATB scores of approximately 19,000 high school students with educational and vocational criteria assessed two years after graduation. Results on predictive validity were promising but inconclusive because of inadequacies in the available criterion data and other methodological problems.

Certain limitations of the GATB should be noted. All tests are highly speeded. Coverage of aptitudes is somewhat limited. No mechanical comprehension test is included, nor are tests of reasoning and inventiveness well represented. The factorial structure of the battery rests on a series of early exploratory studies. A more comprehensive investigation with a large sample and a wider variety of tests would provide a more solid foundation. The consistent reliance on a multiple-cutoff strategy, rather than on regression equations, is also open to question. Although it is probable that certain occupations require critical skills whose lack cannot be compensated by other aptitudes, there is no empirical evidence that this is true of all aptitudes required for all occupations. As for nonlinear relations between aptitudes and job criteria, these too can be handled by regression equations, although the computation of the equation is admittedly more complicated in such cases. The principal justification for the universal use of the multiple-cutoff strategy with the GATB is based on practical limitations in the time and technical training of the personnel interpreting the scores. With the increasing availability of computers, however, this argument loses much of its cogency.

Mention should also be made of the many activities undertaken by the USES to facilitate the testing and job placement of culturally and educationally disadvantaged applicants. Several test orientation procedures have been introduced, including materials and discussion plans designed to allay anxiety and correct misunderstandings about the nature and purpose of testing; booklets giving hints for doing well on tests; and practice exercises with items and answer sheets similar to those used on the GATB. Another innovation was the development of the Basic Occupational Literacy Test (BOLT) for assessing literacy skills of educationally deficient adults. Covering vocabulary, reading comprehension, arithmetic computation, and arithmetic reasoning, this battery is available in four levels. A short, wide-range preliminary test is used to choose the level appropriate for each examinee. The scores on the BOLT are evaluated in terms of occupational literacy requirements rather than school grades.

A major effort to adapt testing procedures to the needs of educationally and culturally disadvantaged adults was the construction of the Nonreading Aptitude Test Battery (NATB) (U.S. Department of Labor, 1970b). This battery was designed to yield the same nine aptitude scores as the GATB, but requires no reading or writing. It includes five of the GATB tests, along with nine tests newly constructed to measure approximately the same functions as the remaining GATB tests. The previously mentioned BOLT screening test is administered to determine whether an examinee has sufficient literacy skills to take the GATB or whether he should take the NATB. On the assumption that the same abilities are measured by both batteries, the NATB utilizes the occupational cutoff scores established for the GATB. The correlations thus far found between corresponding NATB and GATB scores, however, range from the .50's to the .70's and are not high enough to provide adequate justification for this procedure (Steckler, 1973). Administration time for the NATB is quite long, especially for persons unaccustomed to testing. Although the NATB has been in operational use since 1971, it still needs improvement and a project is in progress for this purpose.

The armed services make extensive use of multiple aptitude batteries. These batteries are given for classification purposes after preliminary screening with such instruments as the Armed Forces Qualification Test (AFQT), cited in Chapter 11. Although the Air Force pioneered in the development of classification batteries, all branches of the armed services eventually prepared multifactor batteries for assigning personnel to military occupational specialties. The use of the Aptitude Area scores, derived from the Army Classification Battery, was illustrated in Chapter 7.

In the Air Force, the Aircrew Classification Battery was first developed and employed during World War II for selecting pilots and other flight personnel. An expectancy chart showing pilot stanine scores from this battery and performance in basic flight training was reproduced in Chapter 4 (Fig. 7). Later, an Airman Qualifying Examination was prepared for use with other Air Force Personnel. Both batteries were constructed through factor analysis. More recently, the Airman battery was replaced by the Armed Services Vocational Aptitude Battery, a nine-test classification battery developed jointly by all the armed services (Bayroff & Fuchs, 1970; Vitola & Alley, 1968). Each service combines and uses the basic data from this battery to fit its particular personnel classification needs. The battery is also available to high schools for the testing and counseling of students (*ASVB counselor's manual,* 1974).

On the whole, both the GATB and the various multifactor batteries developed for military use have proved relatively successful as classification instruments. Both differ from the general batteries discussed in the preceding section in that they have been validated chiefly against occupational rather than academic criteria. To be sure, training criteria have sometimes been substituted for actual job performance, but the training programs were job-oriented and quite unlike school work. Both training and job activities of airplane pilots, bakers, beauticians, and the many other kinds of workers included in these testing programs are far removed from traditional academic tasks. With criteria differing more widely from each other and from the verbally loaded academic criteria, there is more room for differential validity.

Another advantage enjoyed by the military batteries is that the number of occupational fields to be covered is smaller than in a general counseling battery. This was especially true of the Aircrew Classification Battery, whose task was limited to the assignment of personnel to four or five occupational specialties. As a result, it was possible to work with relatively narrow group factors, which specifically matched the criterion requirements. In the Air Force research, for example, a large number of different sensorimotor, perceptual, and spatial factors were identified and utilized in test construction. A general counseling battery, on the other hand, must concentrate on a few broad group factors, each of which is common to many jobs. To do otherwise would require the administration of a prohibitive number of tests to each person. But with such broad factors, distinctions are blurred and differential validity drops.

MEASUREMENT OF CREATIVITY

RESEARCH ON CREATIVE TALENT. One of the major developments in psychological testing since midcentury concerns the measurement of creativity. This development is itself only one aspect of an upsurge in research on the nature and cultivation of creative talent (Barron, 1969; Bloomberg, 1973; C. W. Taylor, 1972; Taylor & Barron, 1963; Torrance, 1962; Wallach & Wing, 1969). An increasing number of psychologists and educators have come to recognize that creative talent is not synonymous with academic intelligence and is rarely covered by tests yielding an "IQ." In an early paper, Thurstone (1951) emphasized this distinction and provided a provocative analysis of the possible role of ideational fluency, inductive reasoning, and certain perceptual tendencies in creative behavior. He also called special attention to the contribution of nonintellectual, temperamental factors to creative activity. He observed that creativity is encouraged by a receptive as contrasted to a critical attitude toward novel ideas and that creative solutions are more likely to occur during periods of relaxed, dispersed attention than during periods of active concentration on a problem.

The investigation of creativity has received considerable impetus from the growing demand for research scientists, engineers, and high-level executives. Studies of scientific talent have become increasingly concerned with creative abilities. Interest has shifted from the individual who is merely a cautious, accurate, and critical thinker to the one who also displays ingenuity, originality, and inventiveness. Thus creativity, long regarded as the prime quality in artistic production, is coming more and more to be recognized as a basis for scientific achievement as well.

Investigations of the variables associated with creative achievement have followed a variety of approaches. Some have concentrated on the creative person's biographical history and antecedent experiences; others have analyzed the situational variables conducive to creative productivity. Some investigators have conducted intensive clinical studies of highly eminent scientists. Still others have combined a clinical with a psychometric approach through the use of a variety of personality-testing techniques and controlled observational procedures. This approach is exemplified by the continuing research conducted by MacKinnon (1962) and his associates at the University of California's Institute of Personality Assessment and Research (IPAR). In both the IPAR investigations and other research projects, large and significant differences in a number of personality traits have been found between creative and noncreative groups. Some of the tests employed in this type of research will be examined in Chapter 20.

Standardized tests of creative aptitudes have been produced chiefly in the course of large-scale research projects on the nature of creativity. Two major batteries, to be considered below, are the University of Southern California tests, developed by Guilford and his colleagues, and the Torrance Tests of Creative Thinking. Although commercially available, all these tests are still in experimental form and not ready for operational use.

The items in creativity tests are typically open-ended, thus precluding objective scoring. For this reason, it is imperative to ascertain scorer reliability for all such instruments. Norms are generally tentative, being based on small scattered groups chosen primarily because of availability. Data on reliability and validity vary among individual tests but are usually quite limited. At this stage, the chief application of these tests is in research. The availability of uniform, standardized testing instruments assures a reasonable degree of comparability among different investigations. Eventually some of these tests may be ready for use in clinical, counseling, educational, or vocational situations.

TESTS FROM THE UNIVERSITY OF SOUTHERN CALIFORNIA APTITUDES RESEARCH PROJECT (ARP). Earlier in this chapter, we discussed the factor-analytic research conducted by Guilford and his associates at the University of Southern California. Extending over more than two decades, this research led to the formulation of the structure-of-intellect model (SI). While this model encompasses all intellectual functions, a major contribution of the ARP was in the divergent-production section, about which relatively little prior research was available. In fact, the ARP began primarily as an investigation of reasoning, creativity, and problem-solving.

In the course of his factor-analytic investigations, Guilford developed the categories of divergent and convergent thinking. Convergent thinking leads to a single correct solution determined by the given facts. Divergent thinking, on the other hand, is "the kind that goes off in different directions" (Guilford, 1959, p. 381). Such thinking is less restricted by given facts; it permits changes of direction in problem-solving and leads to a diversity of solutions or products. Figure 65 shows the location of the divergent-production slice of the SI model, together with the individual cells for which tests have thus far been published. These tests are described below, along with the code letters of the principal SI factor measured by each. It will be recalled that this code always gives the Operation, Content, and Product in that order. For example, DSU stands for "divergent production of symbolic units."

SOUTHERN CALIFORNIA TESTS OF FLUENCY, FLEXIBILITY, AND ELABORATION.[5]

Word Fluency (DSU)—write words containing a specified letter. E.g., "O": load, over, pot . . . .

Ideational Fluency (DMU)—name things that belong in the given class. E.g., "Fluids that will burn": gasoline, kerosene, alcohol . . . .

---

[5] The illustrations are parts of sample items, used in explaining instructions in each test. Copyright, Sheridan Psychological Services, and reproduced by permission.

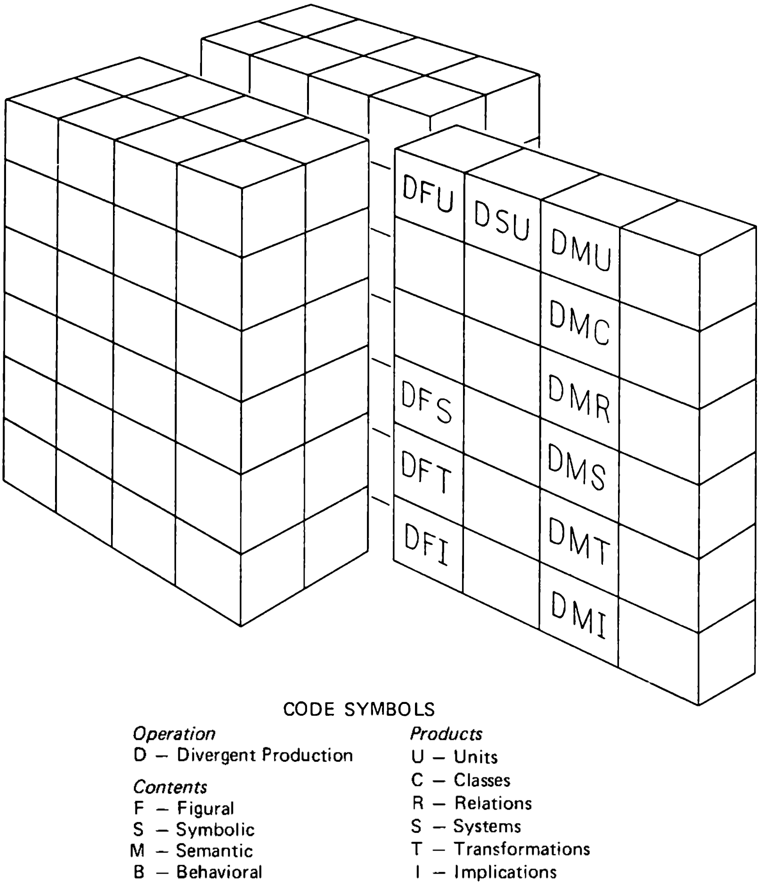

DFU
DSU
DMU
DMC
DMR
DFS
DFT
DMS
DMT
DFI
DMI
CODE SYMBOLS
Operation
D – Divergent Production
Contents
F – Figural
S – Symbolic
M – Semantic
B – Behavioral
Products
U – Units
C – Classes
R – Relations
S – Systems
T – Transformations
I – Implications

Fig. 65. Southern California Tests of Divergent Production Placed within the Structure of Intellect Model.

Associational Fluency (DMR)—write words similar in meaning to the given word. E.g., "Hard": difficult, solid, tough . . . .

Expressional Fluency (DMS)—write four-word sentences, each word to begin with a given letter. E.g., "K——u——y——i——.": Keep up your interest. Kill useless yellow insects.

Alternate Uses (DMC)—list possible uses for a specified object, other than its common use. E.g., "Newspaper (used for reading)": Start a fire. Stuffing to pack boxes . . . .

Simile Interpretations (DMS)—complete in several different ways sentences that contain similes. E.g., "A woman's beauty is like the autumn; it_____.": . . . passes before it can be fully appreciated . . . .

Utility Test (DMU, DMC)—list as many uses as you can for each object. E.g., "Tin Can": A vase for flowers. For cutting round cookies . . . .

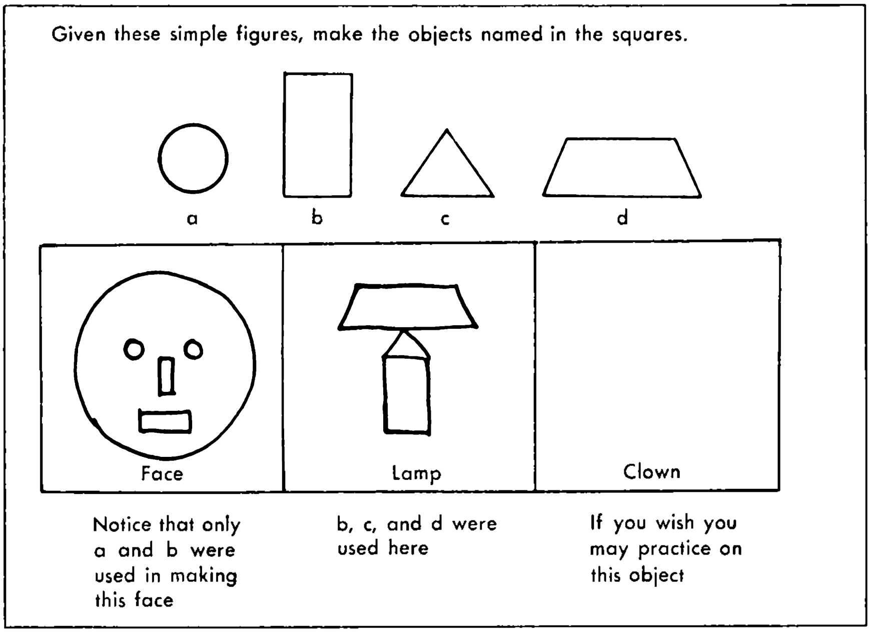
Two scores: total number of responses (ideational fluency—DMU); number of shifts in use categories (spontaneous flexibility—DMC).

Plot Titles (DMU, DMT)—write titles for short-story plots. E.g., "A new clerk in a department store, in anticipation of winter, ordered 10 dozen gloves, but forgot to specify that they should be in pairs. The store now has 100 left-handed gloves.": Southpaw's Delight. Left with a lot of Lefts . . . . Two scores: total number of titles (ideational fluency—DMU); number of "clever" titles, such as those illustrated (originality—DMT).

Consequences (DMU, DMT)—list different consequences of a given hypothetical event. E.g., "What would be the results if people no longer needed or wanted sleep?": Get more work done. Alarm clocks not necessary . . . . Two scores: number of "obvious" responses (ideational fluency—DMU); number of "remote" responses (originality—DMT).

Possible Jobs (DMI)—list possible jobs that might be symbolized by a given emblem. E.g., "Light bulb": electrical engineer, light-bulb manufacturer, a bright student . . . .

Making Objects (DFS)—draw specified objects using only a set of given figures, like circle, triangle, etc. Any given figure may be used repeatedly for the same object and may be changed in size, but no other figures or lines may be added. Figure 66 shows a demonstration item employed as a fore-exercise in this test.



Given these simple figures, make the objects named in the squares.

Notice that only a and b were used in making this face

b, c, and d were used here

If you wish you may practice on this object

Fig. 66. Making Objects: Demonstration Item.

(Reproduced by permission of Sheridan Psychological Services, Inc.)

Sketches (DFU)—each test page contains a set of identical figures, such as circles; make as many different sketches of recognizable objects as possible by elaborating on each figure.

Match Problems (DFT)—remove a specified number of matchsticks, leaving a specified number of squares or triangles. Figure 67 shows part of one of the demonstration items used with this test.

Decorations (DFI)—outline drawings of common objects are to be decorated with as many different designs as possible.

It should be noted that the first ten tests listed above require verbal responses, while the last four employ figural content. These 14 tests represent only some of the divergent production instruments developed in the Aptitudes Research Project (Guilford & Hoepfner, 1971). Other available tests from that project may be published in the future. Still more tests may be added as research continues on the unfilled cells of the SI model.

In the preliminary manuals, Guilford and his co-workers provide tentative norms in terms of *C* scores[6] and percentiles. For most of the tests,

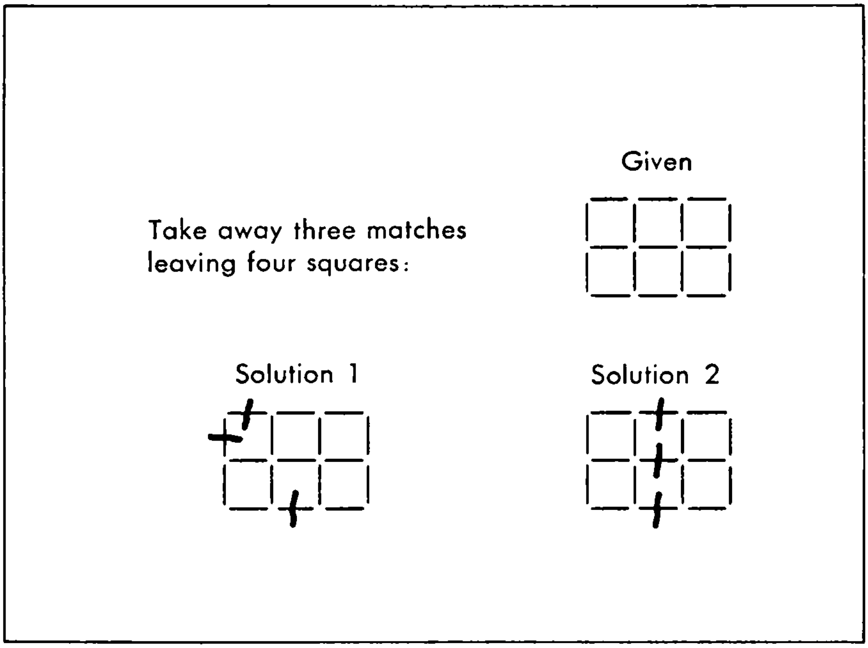


Fig. 67. Match Problems: Part of Demonstration Item.

(Reproduced by permission of Sheridan Psychological Services, Inc.)

---

[6] Similar to stanines, but with 11 units instead of 9.

these norms are given for groups of adults, or ninth-grade students, or both. The tests are generally applicable at the high school level and above.[7] Scorer reliability is not reported in the manuals, but other investigators who have employed these tests in research have obtained coefficients around .90. With a reasonable amount of practice in following the instructions and studying the examples in the manuals, such tests can probably be scored with satisfactory consistency, although the scoring process is quite laborious and time consuming.

Split-half reliability coefficients reported in the manuals run lower than would be desirable, ranging from the .60's to the .80's. For all these tests, factorial validity is reported in terms of the SI factors. Such validity data are based on the extensive research that preceded the publication of each test. Information on criterion-related validity, however, is lacking except for a few scattered independent investigations. A possible limitation of all these tests is their heavy reliance on speed of responses.

TORRANCE TESTS OF CREATIVE THINKING. While the ARP tests were a by-product of factor-analytic research on the nature of intellect, the Torrance tests were developed within an educational context, as part of a long-term research program emphasizing classroom experiences that foster and stimulate creativity (Torrance, 1962, 1963, 1965). Some of the Torrance tests are adaptations of techniques employed in the Southern California tests. Moreover, the scores derived from the complete batteries are based on factors identified in Guilford's research, namely, Fluency, Flexibility, Originality, and Elaboration. In this connection, Torrance explains that he did not seek factorially pure tests. Rather, he tried to contrive situations that provide models of the creative process in its natural complexity. Each test is therefore scored in terms of two, three, or all four of the above factors.

The Torrance Tests of Creative Thinking comprise 12 tests, grouped into a verbal, a pictorial, and an auditory battery. The first battery is labeled *Thinking Creatively with Words*, the second *Thinking Creatively with Pictures*, and the third *Thinking Creatively with Sounds and Words*. In an effort to eliminate threat and to induce a comfortable and stimulating psychological climate, the tests are called "activities" and the instructions place much emphasis on "having fun." The tests are described as suitable from kindergarten to graduate school, although below the fourth grade they have to be administered individually and orally. Two equivalent forms of each battery are available.

In *Thinking Creatively with Words*, the first three activities (Ask-and-

---

[7] A battery of ten creativity tests has also been prepared for children from the fourth grade up; seven of these tests are adaptations of the published adult tests.

Guess) utilize a rather intriguing picture to which the examinee responds by: (1) writing all the questions he would need to ask to find out what is happening; (2) listing possible causes of the action depicted; and (3) listing possible consequences of the action. Activity 4 is concerned with ways of improving a given toy so that children will have more fun playing with it. Activity 5 calls for a list of unusual uses for a common object, as in Guilford's Alternate Uses and Utility tests. Activity 6 requires unusual questions that could be asked about the same object. Activity 7, modeled after Guilford's Consequences, asks for all the things that would happen if a given improbable situation were true. The entire battery yields a total score in each of three traits: Fluency, Flexibility, and Originality.

Thinking Creatively with Pictures consists of three activities. In Picture Construction, a brightly colored curved shape is pasted by the examinee on a blank sheet in a position of his choice and is used as a starting point for drawing an unusual picture "that tells an interesting and exciting story." Picture Completion provides a few lines as a start for drawing a picture in each item. The instructions for this test stress unusual ideas and the scoring is based on aspects of creativity rather than on artistic quality. The last activity provides pairs of short parallel lines (Form A) or circles (Form B), with which as many different pictures as possible are to be produced. Four total scores are obtained from the pictorial battery: Fluency, Flexibility, Originality, and Elaboration.

The most recent addition to the series, *Thinking Creatively with Sounds and Words*, consists of two tests administered by means of long-playing records, which provide both the instructions and the stimuli. The first test, Sounds and Images, utilizes both familiar and unfamiliar sound effects as stimuli. The second, Onomatopoeia and Images, utilizes onomatopoeic words, i.e., words such as crunch or pop that imitate the natural sound associated with the object or action involved. In both tests, the examinee's responses are verbal: after listening to each stimulus, he writes down what it suggests to him. Respondents are urged to give free rein to their imagination. The first test consists of four sound sequences, presented three times; the second consists of ten words, presented four times. Examinees are instructed to give a different association each time. Responses in each test are scored only for Originality, on the basis of empirically established infrequency of occurrence. Thus uncommon responses receive more credit than common responses.

The manuals accompanying the Torrance batteries provide detailed scoring guides, with many examples. The technical manuals cite the results of several studies of scorer reliability, with interscorer correlations falling mostly in the .80's and .90's. In the verbal and pictorial batteries, Originality proved to be the most difficult trait to score, having yielded the only scorer reliabilities under .80. Scattered studies of alternate-form reliabilities over short time intervals, as well as odd-even reliabilities, found coefficients that usually ranged from the .70's to the .90's. In general, the verbal scores show higher reliabilities than the figural scores. Tentative norms for each of the three verbal, four figural, and two auditory scores are given as *T* scores in terms of available groups of schoolchildren and adults. Means and *SD*'s of several other groups are also included for comparative purposes.

As in the Southern California tests, speed is an integral part of performance on the Torrance tests, a characteristic common to virtually all current creativity tests. The technical manuals summarize a number of investigations in which the Torrance tests were employed for a variety of research purposes. All these unrelated studies contribute toward the construct validation of the tests, but no clear picture emerges as yet. Most of the studies were conducted on schoolchildren, although some data on high school and college students and on occupational adult groups are also cited. Many suggestive relationships have been found between Torrance scores and interests, attitudes, and other personality traits measured by tests, ratings, or other assessment procedures. Significant improvement in Torrance scores has also been observed as a result of educational experiences designed to stimulate creative thinking, but no external evidence is provided to indicate whether such improvement extends to other, nontest situations.

In general, evidence of relation between the Torrance tests and everyday-life criteria of creative achievement is meager. There is need for comprehensive, systematic validation studies showing the relation of these tests to measures of other intellectual traits as well as to practical criteria. The use of Fluency, Flexibility, Originality, and Elaboration scores that cut across the individual tests is questionable. A factor analysis of the performance of about 800 fifth-grade children on the Torrance tests gave no support to the interpretation of these scores in terms of single constructs (Yamamoto & Frengel, 1966). On the contrary, the factors identified were task-specific. The intercorrelations of different scores derived from a single test were higher than the intercorrelations of similarly labeled scores (e.g., Fluency) derived from different tests. The two Originality scores obtained from the third battery yield correlations in the .30's and .40's with each other and still lower correlations with Originality scores from the verbal and figural batteries. It would thus seem inadvisable to derive more than one score from any one test or to treat similarly labeled scores from different tests as measures of a single trait.

CREATIVE ACHIEVEMENT. Although tests of divergent production, such as those of Guilford and Torrance, probably come close to measuring the essential aspects of creativity, other abilities are undoubtedly needed for effective creative achievement, especially in the sciences. In their zeal to counteract the earlier concentration on comprehension and memory, as represented by traditional intelligence tests, some investigators may have swung too far in the opposite direction. For genuine creative achievement, the uninhibited divergent-production phase must eventually be followed by critical evaluation. In the technique popularly known as "brainstorming," for example, creativity is stimulated by the temporal separation of the productive and evaluative phases of creative activity. A critical, evaluative attitude at an early stage of creative production may seriously thwart the development of new ideas. But critical evaluation is to be only temporarily deferred, not permanently abolished.

Several cognitive and evaluative aptitudes usually classified under reasoning have emerged from the Guilford research. Some of the published tests from the Aptitudes Research Project measure these factors. Among them are the following: Ship Destination Test, which is heavily loaded with a general reasoning factor and utilizes problems similar to those found in arithmetic reasoning tests; Logical Reasoning, composed of items in syllogistic form; and Pertinent Questions, designed as a measure of conceptual foresight. Another test covering several aspects of effective reasoning is the Watson-Glaser Critical Thinking Appraisal. Designed for high school and college levels, this test contains five parts, dealing with inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments.

In discussing creative productivity in the arts, Guilford (1957, 1967) suggests that several of the divergent-production factors so far identified may play an important part. Those in the verbal area, with which many of the available tests are concerned, are probably related to creative writing. Corresponding factors pertaining to visual, auditory, or even kinaesthetic figural content—many of which have not yet been identified—may play an equally important part in the graphic arts, music, and choreography. In addition, creative productivity in the arts, as in the sciences, undoubtedly requires a certain minimum proficiency in relevant comprehension and memory factors, such as verbal comprehension, spatial orientation, visual or auditory memory, and the like.

It is thus apparent that creative achievement—whether in science, engineering, art, music, or other fields of human endeavor—requires a complex pattern of aptitudes and personality traits appropriate to the particular field. Current tests of creativity concentrate on certain elements of this pattern that had heretofore been largely neglected in psychometrics. But they are not intended to replace other types of available tests.

CHAPTER 14

Educational Testing

NEARLY every type of available test is utilized in the schools. Intelligence, special aptitude, multiple aptitude, and personality tests can all be found in the repertory of the educational counselor and the school psychologist. Teachers and educational administrators frequently have to act on the results obtained with several different kinds of tests. Certain types of tests, however, have been specifically developed for use in educational contexts, predominantly at the elementary and high school levels. It is with these tests that the present chapter is concerned. They include both predictive instruments designed for specific educational purposes and many varieties of educational achievement tests.

ACHIEVEMENT TESTS: THEIR NATURE AND USES

NATURE. Surpassing all other types of standardized tests in sheer numbers, achievement tests are designed to measure the effects of a specific program of instruction or training. It has been customary to contrast achievement tests with aptitude tests, the latter including general intelligence tests, multiple aptitude batteries, and special aptitude tests. From one point of view, the difference between achievement and aptitude testing is a difference in the degree of uniformity of relevant antecedent experience. Thus, achievement tests measure the effects of relatively standardized sets of experiences, such as a course in elementary French, trigonometry, or computer programming. In contrast, aptitude test performance reflects the cumulative influence of a multiplicity of experiences in daily living. We might say that aptitude tests measure the effects of learning under relatively uncontrolled and unknown conditions, while achievement tests measure the effects of learning that occurred under partially known and controlled conditions.

A second distinction between aptitude and achievement tests pertains to their respective uses. Aptitude tests serve to predict subsequent performance. They are employed to estimate the extent to which the individual will profit from a specified course of training, or to forecast the quality of his achievement in a new situation. Achievement tests, on the other hand, generally represent a terminal evaluation of the individual's status on the completion of training. The emphasis in such tests is on what the individual can do at the time. This difference is perhaps most clearly illustrated by the procedures for estimating the validity of achievement tests, as contrasted with those followed in validating aptitude tests. Although predictive criterion-oriented validity is the most direct way of assessing aptitude tests, achievement tests are characteristically evaluated in terms of their content validity (Ch. 6).

It should be recognized, however, that no distinction between aptitude and achievement tests can be rigidly applied. Some aptitude tests may depend on fairly specific and uniform prior learning, while some achievement tests cover relatively broad and unstandardized educational experiences. Similarly, an achievement test may be used as a predictor of future learning. As such it serves the same purpose as an aptitude test. For example, the progress a pupil has made in arithmetic, as determined by his present achievement test score, may be employed to predict his subsequent success in algebra. Achievement tests on premedical courses can serve as predictors of performance in medical school. Whenever different individuals have had the same or closely similar courses of study, achievement tests based on such courses may provide efficient indices of future performance. When used for predictive purposes, of course, achievement tests should be evaluated in terms of their criterion correlations.

In differentiating between aptitude and achievement tests, we should especially guard against the naïve assumption that achievement tests measure the effects of learning, while aptitude tests measure "innate capacity" independent of learning. This misconception was fairly prevalent in the early days of psychological testing, but has been largely corrected in the subsequent clarification of psychometric concepts. It should be obvious that all psychological tests measure the individual's current behavior, which inevitably reflects the influence of prior learning. The fact that every test score has a "past" does not, however, preclude its having a "future." While revealing the effects of past learning, test scores may, under certain conditions, serve as predictors of future learning.[1]

A useful concept that is coming to replace the traditional categories of aptitude and achievement in psychometrics is that of *developed abilities*. All ability tests—whether they be designed as general intelligence tests, multiple aptitude batteries, special aptitude tests, or achievement tests—measure the level of development attained by the individual in one or more abilities. No test reveals how or why the individual reached that level. Existing tests of developed abilities may be ordered along a continuum in terms of the specificity of experiential background that they presuppose. This continuum is illustrated in Figure 68. At one extreme are the course-oriented achievement tests covering narrowly defined technical skills or factual information. A test in Russian vocabulary or television maintenance would fall at this end of the continuum. Many of the achievement tests employed for personnel selection or classification in industry or the armed services are of this type. Teacher-made classroom examinations also belong in this category.

---

[1] For a further discussion of the distinction between aptitude and achievement tests, from many viewpoints, see the conference report edited by D. R. Green (1974), especially the chapters by Humphreys (Ch. 8) and Ebel (Ch. 10), and the review of this report by Anastasi (1975).

Next in the continuum come the broadly oriented achievement tests commonly used today to assess the attainment of major, long-term educational goals. Here we find tests focusing on the understanding and application of scientific principles, the interpretation of literature, or the appreciation of art. Still broader in orientation are tests of intellectual skills that affect the individual's performance in a wide variety of activities. These skills may themselves be rather narrowly defined, as in certain common work-study skills (e.g., interpretation of tables, use of indexes and dictionaries). Or they may be broader, such as reading comprehension and arithmetic computation. At the broadest level, we find achievement tests designed to measure the effects of education on logical thinking, critical evaluation of conclusions, problem-solving techniques, and imagination. It is apparent that at this point achievement tests fuse imperceptibly with traditional intelligence and aptitude tests. This overlapping of "aptitude" and "achievement" tests can easily be demonstrated empirically. An examination of the content of several current instruments classified as intelligence and as achievement tests, respectively, reveals close similarity of content. It has long been known, moreover,

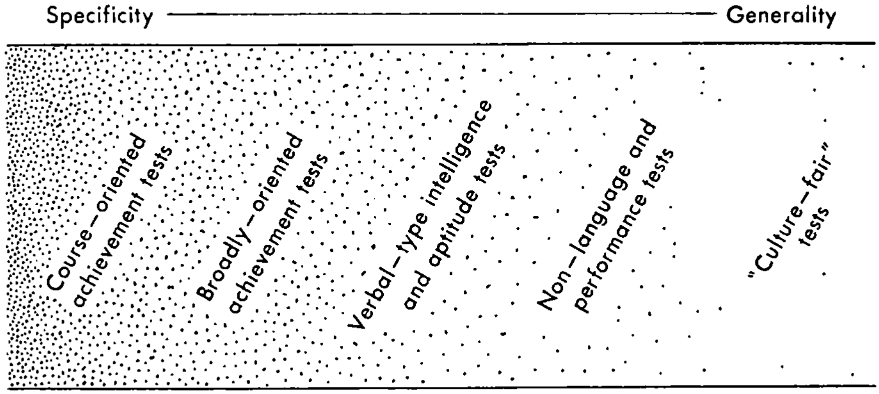

Specificity
Generality
Course-oriented achievement tests
Broadly-oriented achievement tests
Verbal-type intelligence and aptitude tests
Non-language and performance tests
"Culture-fair" tests

FIG. 68. Tests of Developed Abilities: Continuum of Experiential Specificity.

that intelligence tests correlate about as highly with achievement tests as different intelligence tests correlate with each other (Coleman & Cureton, 1954; Kelley, 1927, pp. 193–209). In some instances, in fact, the correlation between achievement and intelligence tests is as high as the reliability coefficients of each test.

Further reference to Figure 68 shows a wide disparity among traditional intelligence tests with regard to the experiential or educational specificity of their content. The predominantly verbal intelligence or aptitude tests, such as the Stanford-Binet or the Differential Aptitude Tests, are closest to the achievement tests. Next come nonlanguage and performance tests, and finally the "culture-fair" tests designed for use with persons of widely varied experiential backgrounds (Ch. 10). As a safeguard against possible misinterpretation of either aptitude or achievement test scores, this schematic continuum should be kept constantly in mind.

Labeling some instruments as "aptitude tests" and others as "achievement tests" has led to certain misuses of test results. A common example is the practice of identifying as underachievers those children whose achievement test scores are lower than their scholastic aptitude or intelligence test scores. Actually, such intraindividual differences in test scores reflect the universal fact that no two tests (or other performance indicators such as course grades) correlate perfectly with each other. The question of under- or overachievement can be more accurately formulated as overprediction or underprediction from the first to the second test. Among the reasons for the prediction errors in individual cases are the unreliability of the measuring instruments, differences in content coverage, the varied effects of attitudinal and motivational factors on the two measures, and the impact of such intervening experiences as remedial instruction or a long illness.[2]

It should be noted that underprediction and overprediction will occur regardless of the type of test used, i.e., if an intelligence test is used to predict subsequent achievement test performance, or if an achievement test is used to predict subsequent intelligence test performance; and it will occur in estimating scores on the later test from scores on the earlier test or vice versa. Finally, the administration of alternate forms or different levels of an achievement test before and after a course of instruction permits a more accurate analysis of individual accomplishment than does the use of two different tests. To take an extreme example, if achievement in reading comprehension is predicted from a nonverbal intelligence test that is heavily loaded with spatial aptitude, the children with higher spatial than verbal aptitude will look like underachievers, while those with higher verbal than spatial aptitude will look like overachievers.

---

[2] For a comprehensive analysis of methodological problems in the measurement of underachievement, see Thorndike (1963).

USES. The many roles that achievement tests can play in the educational process have long been recognized. As an aid in the assignment of grades, such tests have the advantages of objectivity and uniformity. If properly constructed, they have other merits, such as adequacy of content coverage and reduction of the operation of irrelevant and chance factors in marking procedures. Achievement tests also constitute an important feature of remedial teaching programs. In this connection, they are useful both in the identification of pupils with special educational disabilities and in the measurement of progress in the course of remedial work.

For all types of learners, the periodic administration of well-constructed and properly chosen achievement tests serves to facilitate learning. Such tests reveal weaknesses in past learning, give direction to subsequent learning, and motivate the learner. The incentive value of "knowledge of results" has been repeatedly demonstrated by psychological experiments in many types of learning situations, with subjects of widely varying age and education. The effectiveness of such self-checking is generally heightened by immediacy. Thus, when achievement examinations are employed primarily as a learning aid, it is desirable for the students to become aware of their errors as soon after taking the tests as possible.

From another angle, achievement tests provide a means of adapting instruction to individual needs. Teaching can be most fruitful when it meets the learner at whatever stage he happens to be. Ascertaining what each individual is already able to do and what he already knows about a subject is thus a necessary first step for effective teaching. The growth of fall testing programs points up the increasing use of test results as a basis for planning what is to be taught to a class as a whole and what modifications and adjustments need to be made in individual cases. By giving tests at the beginning of the school year, constructive steps can be taken to fill the major gaps in knowledge revealed by the test results. Further examples of the role of achievement tests in the teaching process can be found in connection with criterion-referenced testing, individually tailored instructional systems, and computer-aided learning procedures, discussed in Chapter 4.

Finally, achievement tests may be employed as aids in the evaluation and improvement of teaching and in the formulation of educational goals. Achievement tests can provide information on the adequacy with which essential content and skills are actually being taught. They can likewise indicate how much of the course content is retained and for how long. Are certain types of material retained longer than others? What are the most common errors and misunderstandings encountered? How well can the learners apply their knowledge to new situations? By focusing attention on such questions and by providing concrete facts, achievement tests stimulate an analysis of training objectives and encourage a critical examination of the content and methods of instruction. Public demands for educational accountability require the proper use of well-constructed achievement tests to assess the results of the educational process (Gronlund, 1974; Wrightstone, Hogan, & Abbott, 1972). It is especially important for this purpose to recognize the many conditions that affect both test performance and school learning, including the cumulative effects of antecedent and continuing experiences outside of the school environment.

GENERAL ACHIEVEMENT BATTERIES

NATURE AND SCOPE. Several batteries are available for measuring general educational achievement in the areas most commonly covered by academic curricula. This type of test can be used from the primary grades to the adult level, although its major application has been in the elementary school. Typically these batteries provide profiles of scores on individual subtests or in major academic areas. An advantage of such batteries as against independently constructed achievement tests is that they may permit horizontal or vertical comparisons, or both. Thus, an individual's relative standing in different subject-matter areas or educational skills can be compared in terms of a uniform normative sample. Or the child's progress from grade to grade can be reported in terms of a single score scale. The test user should check whether a particular battery was so standardized as to yield either or both kinds of comparability.

To provide an overview of the nature and grade range of general achievement batteries, some representative examples have been listed in Table 29, together with the grade levels covered by each. It will be noted that some of these batteries are designed for the elementary grades, others for the high school; but most span a broad range extending into both levels and occasionally even into college. Although a few provide a single battery for the grade range covered, the large majority have several overlapping batteries for use at different grade levels. Some separately listed batteries for the elementary and high school levels actually form a coordinated series, permitting comparable measurement from Grades 1 through 12. One such combination is the Iowa Tests of Basic Skills and the Tests of Academic Progress; another is the Stanford Achievement Test and the Stanford Test of Academic Skills.

Another major difference among available batteries pertains to their relative emphasis on educational skills and knowledge of content. A primary focus on skills is characteristic of the California Achievement Tests,

TABLE 29
Representative Achievement Batteries

	Grade Range														
Battery	K	1	2	3	4	5	6	7	8	9	10	11	12	13	14
California Achievement Tests: 1970 Edition		x	x	x	x	x	x	x	x	x	x	x	x		
Iowa Tests of Basic Skills		x	x	x	x	x	x	x	x						
Iowa Tests of Educational Development										x	x	x	x		
Metropolitan Achievement Test: 1970 Edition	x	x	x	x	x	x	x	x	x	x					
SRA Achievement Series		x	x	x	x	x	x	x	x	x					
Sequential Tests of Educational Progress (STEP) —Series II					x	x	x	x	x	x	x	x	x	x	x
Stanford Achievement Test[a]		x	x	x	x	x	x	x	x	x					
Stanford Test of Academic Skills (TASK)									x	x	x	x	x	x	
Tests of Academic Progress (TAP)										x	x	x	x		

[a] The Stanford Early School Achievement Test (SESAT), to be discussed later in this chapter, extends this series to the beginning of kindergarten.

Iowa Tests of Basic Skills, Iowa Tests of Educational Development, and SRA Achievement Series. These tests concentrate on such educational skills as reading, arithmetic, spelling and language usage, and work-study skills (e.g., map reading, reading graphs, use of references). When achievement in different subject-matter areas is to be measured—particularly at the high school level where specialized courses are common—this type of test may concentrate on the student's ability to understand passages from each content area. The Iowa Tests of Educational Development, for example, include tests on the ability to interpret reading materials in literature, the social studies, and the natural sciences. Some batteries emphasize the application of knowledge and learned skills to the solution of new problems. A notable example of this approach is provided by the Sequential Tests of Educational Progress (STEP)—Series II.

STANFORD ACHIEVEMENT TEST. As a specific example of an achievement battery, we may consider the Stanford Achievement Test (SAT). First published in 1923, this pioneer battery has undergone extensive changes in successive editions, the latest appearing in 1973. The SAT spans grades one to nine, in six batteries. Most of these batteries require approximately four to five hours of testing time, distributed over several sessions. The tests are also available in shorter "basic" batteries, which omit certain specialized content areas. The reading and the mathematics subtests may be separately administered. With few exceptions, the same subtests occur at the different levels. All batteries are available in two parallel forms. There is a practice booklet, with test-taking suggestions and typical items, which may be administered a few days before the test proper.

To illustrate the content of the SAT, we may examine Intermediate Level II, applicable from the middle of the fifth to the end of the sixth grade. This battery comprises the following 11 subtests:

1. *Vocabulary:* to measure vocabulary development independently of reading ability, incomplete sentences are presented orally; child chooses correct word to complete each sentence. E.g., A person who is in a bad mood most of the time is called a ________. (1) hermit (2) comedian (3) grouch (4) student.

2. *Reading Comprehension:* prose passages and poem read by child, each followed by a set of questions; requires such skills as identifying central idea of passage, drawing inferences, recognizing implicit meaning, and responding to contextual cues.

3. *Word Study Skills:* identifying sounds of letters and letter combinations, presented visually; combining syllables to make a word.

4. *Mathematics Concepts:* understanding of mathematical terms, notation, and operations; fractions, sets, percentages, geometric concepts, etc.

5. *Mathematics Computation:* involving a variety of operations, all items are presented entirely in numerical terms.

6. *Mathematics Applications:* verbally presented problems, sampling everyday-life situations; includes typical arithmetic reasoning problems, as well as items involving measurement, graphs, scaled drawings, etc. (see Fig. 69).

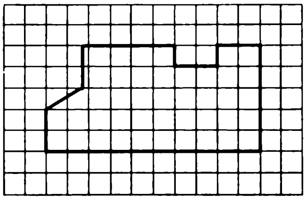
7. *Spelling:* identifying incorrectly spelled words as well as use of wrong homophones in context (e.g., wood for would).

8. *Language:* correct language usage, including capitalization, punctuation, verb and pronoun forms, and sentence structure.

9. *Social Science:* designed to test for factual information as well as high-order reasoning skills; draws from geography, history, economics, political science, anthropology, and sociology (see Fig. 70).

MATHEMATICS APPLICATIONS

Our class will make a big map from this small map of our town. We have enough room on the board to make it 4 feet high. How many feet long will it be? (If the correct answer is *not here,* mark NH.)



a 7 b 5 c 10 d 8 e NH

LISTENING COMPREHENSION

Examiner reads passage, questions, and response options:

Questions 21 through 24 are based on this passage. Listen carefully.

Grass should not be cut too short. It should be about three inches high. One reason is because a properly high-cut lawn acts as its own crabgrass killer. Crabgrass is less than an inch high and can't grow if it's shaded by higher grass. Another reason is that a high cut shades the ground and keeps the roots cool and moist.

21. The best title for this passage would be —
    1. How to Mow a Lawn
    2. How to Grow Crabgrass
    3. How to Make a Garden
    4. How to Use Fertilizer

22. Crabgrass is apparently —
    1. highly valued
    2. very tall
    3. poisonous
    4. unpopular

23. What word best describes this passage?
    1. reasoning
    2. advice
    3. humor
    4. mystery

24. Grass roots should be kept —
    1. warm and dry
    2. warm and moist
    3. cool and moist
    4. cool and dry

FIG. 69. Typical Items from Two Subtests of Stanford Achievement Test, Intermediate Level II, for Grades 5 and 6.

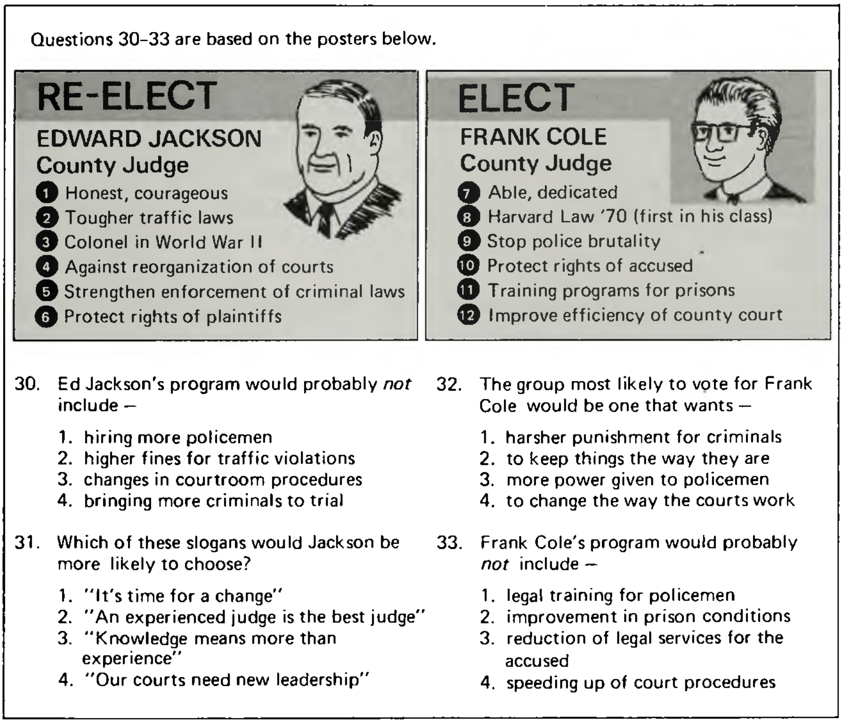
(Reproduced by courtesy of Harcourt Brace Jovanovich, Inc.)

10. *Science:* covers scientific method as well as understanding of basic conconcepts in physical and biological sciences.

11. *Listening Comprehension:* immediately after listening to each passage read aloud by examiner, child answers a set of questions about it; questions are also presented orally (see Fig. 69).

All items are of the multiple-choice form. Several tests, such as Vocabulary and Reading Comprehension, cover a wide diversity of content, chosen to assess information from many fields and to appeal to the interests of children from various cultural backgrounds. In line with current interest in objective-referenced (criterion-referenced) interpretations, an Index of Instructional Objectives has been prepared for each level, stating specifically what each item or item group was designed to test. Such descriptions serve the dual function of providing information about the content validity of the tests and permitting an analysis of each child's performance in terms of specific skills. The interpretive manual for each level lists somewhat broader instructional objectives corresponding to item groups and includes suggestions for teaching any skills found to be deficient in an individual or a class.

Norms are available for each of the 11 subtests, as well as for certain area totals, including total reading, total mathematics, and total auditory scores (combined scores on the aural tests of Vocabulary and Listening Comprehension), and for the total battery. Scores can be interpreted in terms of grade norms and in terms of percentiles and stanines within



Questions 30–33 are based on the posters below.

RE-ELECT
EDWARD JACKSON
County Judge
① Honest, courageous
② Tougher traffic laws
③ Colonel in World War II
④ Against reorganization of courts
⑤ Strengthen enforcement of criminal laws
⑥ Protect rights of plaintiffs

ELECT
FRANK COLE
County Judge
⑦ Able, dedicated
⑧ Harvard Law '70 (first in his class)
⑨ Stop police brutality
⑩ Protect rights of accused
⑪ Training programs for prisons
⑫ Improve efficiency of county court

30. Ed Jackson's program would probably *not* include –
    1. hiring more policemen
    2. higher fines for traffic violations
    3. changes in courtroom procedures
    4. bringing more criminals to trial

31. Which of these slogans would Jackson be more likely to choose?
    1. "It's time for a change"
    2. "An experienced judge is the best judge"
    3. "Knowledge means more than experience"
    4. "Our courts need new leadership"

32. The group most likely to vote for Frank Cole would be one that wants –
    1. harsher punishment for criminals
    2. to keep things the way they are
    3. more power given to policemen
    4. to change the way the courts work

33. Frank Cole's program would probably *not* include –
    1. legal training for policemen
    2. improvement in prison conditions
    3. reduction of legal services for the accused
    4. speeding up of court procedures

Fig. 70. Typical Items from Social Science Subtest of Stanford Achievement Test, Intermediate Level II, for Grades 5 and 6.

(Reproduced by courtesy of Harcourt Brace Jovanovich, Inc.)

grades. In addition, all scores can be expressed as scaled scores on a scale that is continuous through all levels of the SAT. The same score scale ties SAT with the Stanford Early School Achievement Test (SESAT) and the Stanford Test of Academic Skills (TASK), thereby providing continuous measurement from kindergarten through high school and into community college. It should be noted, however, that these scaled scores are not comparable across the different subtests.

The 1973 SAT was standardized on a national sample of the school population in grades one to nine, including over 275,000 pupils and stratified in accordance with the 1970 U.S. Census with regard to proportion of principal ethnic minorities; geographic region; size, income, and educational level of community; and major school characteristics. Percentiles and stanines were established empirically by administering the tests at the beginning, middle, and end of the school year in grades one and two, and at the beginning and end of the school year in all other grades. The time of year when tests are given is accordingly taken into account when converting scores to within-grade percentiles or stanines.

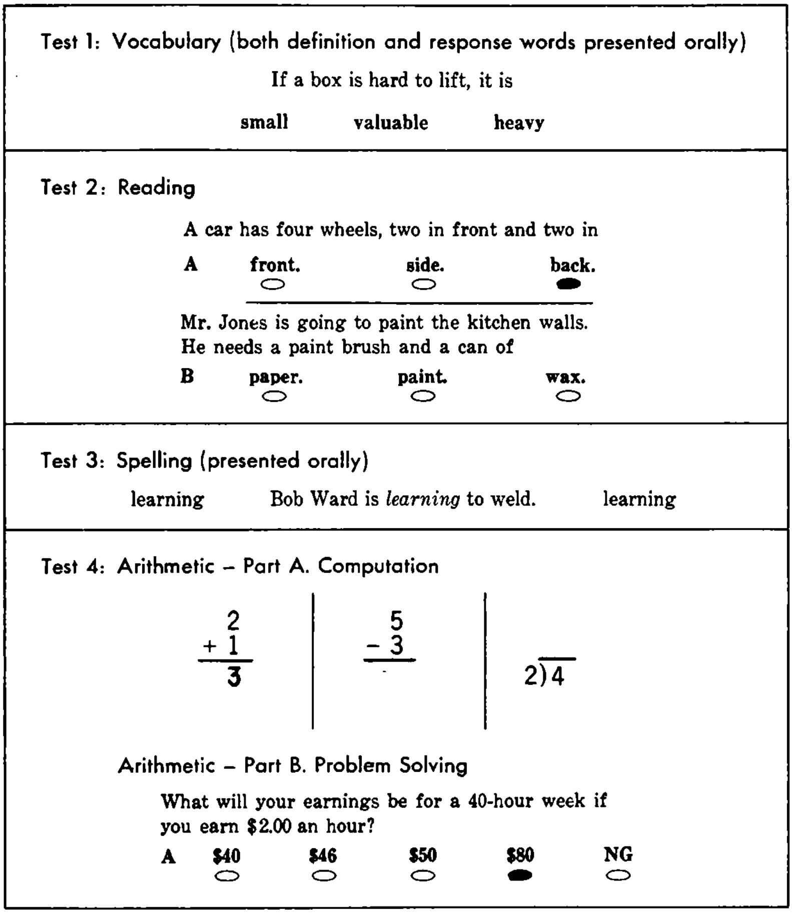
The reliabilities of SAT subtests and total scores are consistently high. The test authors warn, however, that the reliabilities of item groups measuring a specific instructional objective or skill will necessarily be considerably lower. Performance on item groups can be more confidently interpreted for class means than for individuals. With regard to content validity, the development of the 1973 SAT demonstrates outstanding technical quality. Unusual ingenuity is evidenced in the design of objective items to measure a wide variety of complex cognitive functions.

TESTS OF BASIC EDUCATION FOR ADULTS. The increasing concern for culturally and economically disadvantaged minorities has stimulated the development of new tests, not only for preschool and primary-school children, but also for undereducated adults. An example of the latter type of test is the Adult Basic Learning Examination (ABLE). This is an achievement test specifically designed for use in connection with adult-education classes, educational programs conducted in penal institutions, and job training programs. It is available in three levels corresponding to Grades 1–4 (Level I), 5–8 (Level II), and 9–12 (Level III). At each level, there are two equivalent forms. Drawing largely from practical problems of adult daily life, ABLE consists of the following four tests, illustrated in Figure 71:

1. *Vocabulary:* examinee chooses correct word to fit orally presented definitions.
2. *Reading:* paragraph comprehension, inference, and associated general information.

3. *Spelling:* writing words from dictation.
4. *Arithmetic:* number computation and problem-solving—requires no reading at Level I.

Scores on each of the four tests may be evaluated in terms of grade equivalents, established by administering both ABLE and the Stanford Achievement Test to samples of elementary and high school students. Percentile and stanine norms are also available from large groups enrolled in adult basic education programs and from military samples. Split-half and Kuder-Richardson reliabilities of each test in adult groups range from .80 to .96. Correlations between corresponding tests of ABLE



Test 1: Vocabulary (both definition and response words presented orally)

If a box is hard to lift, it is

small **valuable** **heavy**

Test 2: Reading

A car has four wheels, two in front and two in

A **front.** ○ **side.** ○ **back.** ●

Mr. Jones is going to paint the kitchen walls.
He needs a paint brush and a can of

B **paper.** ○ **paint.** ○ **wax.** ○

Test 3: Spelling (presented orally)

learning Bob Ward is *learning* to weld. learning

Test 4: Arithmetic – Part A. Computation

$$\begin{array}{r} 2 \\ +\,1 \\ \hline 3 \end{array} \qquad \begin{array}{r} 5 \\ -\,3 \\ \hline \end{array} \qquad 2\overline{)4}$$

Arithmetic – Part B. Problem Solving

What will your earnings be for a 40-hour week if you earn $2.00 an hour?

A **$40** ○ **$46** ○ **$50** ○ **$80** ● **NG** ○

Fig. 71. Sample Items from Adult Basic Learning Examination (ABLE), Level II.

(Reproduced by permission of Harcourt Brace Jovanovich, Inc.)

and the Stanford Achievement Test in the elementary and high school samples range from the .60's to the .80's. Correlations of the same general magnitude were also found in several adult samples. In a study of approximately 800 military recruits, scores on ABLE Level III were checked against performance on the tests of General Educational Development (GED) (Waggener, Starr, & Hopkins, 1974). The latter tests are employed in establishing high school equivalency for education received in the armed services. The correlations between ABLE tests and the most nearly comparable GED tests ranged from the high .50's to the .70's. Multiple correlations for predicting each GED score from ABLE tests ranged from .67 to .75. A similar study with ABLE Level II yielded lower but still substantial correlations.

Another test designed for the same general purpose as ABLE is the USES Basic Occupational Literacy Test (BOLT) described in Chapter 13. Nearly all major (and some minor) test publishers have prepared similar tests, some of which are listed in Appendix E. Most are crude instruments, which reveal their origins in crash programs.

At a higher educational level, similar tests have been developed for use in the remedial training of college students with deficiencies in basic educational skills. Such remedial programs have become especially important following the introduction of open admissions policies by a number of colleges. Tests are needed both to identify students requiring remedial training and to assess their progress during training. An example of this type of test is the McGraw-Hill Basic Skills System, which includes both tests and instructional materials. Separate sections cover reading, writing, spelling, vocabulary, study skills, and mathematics.

STANDARDIZED TESTS IN SEPARATE SUBJECTS

While the general achievement batteries have been focusing more and more on basic educational skills, there remains a need for well-constructed instruments to measure achievement in the many specialized areas included in educational curricula. This need increases progressively from the lower to the higher levels of the educational system, as students branch out from the common educational core. It should be noted, too, that nearly all achievement batteries make available separately printed parts covering particular areas. The most common are achievement tests in reading and in mathematics; but tests covering other broad content areas can also be obtained.

Separate standardized tests have been prepared for almost every subject, from American history to physical education, and from accounting to physics. A comprehensive introduction to achievement tests in different instructional fields can be found in Gerberich, Greene, and Jorgensen (1962, Chs. 15–24). The *Mental Measurements Yearbooks* also cover this type of test quite fully, under the appropriate subject-matter categories. A noteworthy development since the 1960's has been the publication of tests in modern foreign languages that use tape recordings to test proficiency in reading, writing, listening, and speaking. Examples of these tests and of recently developed tests in other instructional areas are listed in Appendix E.

Of particular interest is the coordinated series of achievement tests for different high school subjects included in the annual testing program of the College Entrance Examination Board (1974a). A major feature of such a coordinated series is the provision of a single system of comparable scores for all tests. It is thus possible to make direct comparisons among scores obtained in different subject-matter areas. Unlike the achievement batteries discussed earlier in this chapter, however, these tests cannot be standardized on a single normative population. Moreover, it is likely that the standardization samples available for the various academic subjects differ appreciably in general scholastic aptitude. For example, students taking an examination in advanced mathematics are probably more highly selected than those taking an examination in American history. If scores on all these tests are evaluated with reference to the means of their respective normative samples, an individual might, for instance, appear to be more proficient in American history than in mathematics, simply because he was compared with an academically weaker normative sample in the former case.

The College Board tests employ a common standard score scale that is adjusted for the differences among normative samples for each test (Angoff, 1971b). The scores on the Scholastic Aptitude Test, taken by all candidates, provide the basis for such adjustments. All raw scores on achievement tests are converted into a standard score scale having a mean of 500 and an *SD* of 100 in the fixed reference group. As explained in Chapter 4, this group consists of the 10,651 subjects who took the SAT in 1941. The scores obtained by the candidates taking any achievement test during subsequent years are expressed with reference to this group. The adjustment takes into account any differences in the distribution of SAT scores between the group taking an achievement test in a particular subject and the fixed reference group.

The availability of comparable achievement test scores across fields is desirable for many purposes. It is clear that some adjustment must be made to allow for the differences among the examinee populations in different fields. A word of caution is in order, however, regarding the interpretation of the system of scaled scores described above. Close comparability of scores can be attained only when the norming test and the anchor test correlate very highly with each other. It cannot be assumed that the relative status of different groups on a scholastic aptitude or general intelligence test is necessarily the same as their relative status on other tests. For example, if the candidates taking a physics test and those taking a Latin test were to obtain the same distribution of scholastic aptitude scores, they might still differ significantly in their aptitudes for Latin and for physics. In other words, we cannot assume that the Latin sample would have obtained the same scores in physics as the present physics sample, if both had pursued the same courses.

To be sure, such an assumption is not implied by the use of the above system of scaled scores. This system merely expresses all scores in terms of a fixed standard that can be precisely and operationally defined. But the test user with only a superficial knowledge of how the scores were derived could easily be misled into unwarranted interpretations. It must be borne in mind that the procedures followed in developing such a scoring system provide comparability of a sort. But they do not necessarily yield the identical norms that would be obtained if all students had been enrolled in the same courses of study and had been given all the tests in each series.

We may inquire where the type of achievement test discussed in this section fits into the total testing picture. Such tests are obviously well suited for use as end-of-course examinations. But they may serve other functions also. In comparison with the more broadly oriented tests of educational development, traditional achievement tests, which are more closely linked to specific courses, measure more nearly distinct skills and knowledge. For this reason, they are likely to yield lower correlations with intelligence tests than have been found for broad achievement tests. If combined with intelligence tests, therefore, the specialized achievement tests will contribute more unique, nonoverlapping variance and may permit better prediction of subsequent outcomes.

There is also an increasing utilization of specialized achievement tests as college equivalency examinations. High school students with additional preparation in certain areas may take tests in the College Board's Advanced Placement Program (APP) to gain admission to college with advanced standing in one or more subjects. A related development is to be found in the College-Level Examination Program (CLEP), also administered by the College Board. The general purpose of this program is to facilitate the granting of college credit by examination and to provide a national system for evaluating college-level education acquired through independent study and other nontraditional procedures.

TEACHER-MADE CLASSROOM TESTS

Undoubtedly the largest number of tests covering the content of specific courses or parts of courses are prepared by instructors for use in their own classrooms. The vast diversity among courses on the same subject and with identical titles—particularly at the high school level and beyond—is well known. Under these conditions, no external standardized test could suffice. The preparation of local classroom tests, however, can be substantially improved by an application of the techniques and accumulated experience of professional test constructors. The development of classroom tests can be divided into three major steps: (1) planning the test, (2) item writing, and (3) item analysis. Some simple techniques of item analysis, suitable for use with small groups, were described in Chapter 7. An overview of the other two steps is given below.

The test constructor who plunges directly into item writing is likely to produce a lopsided test. Without an advance plan, some areas will be overrepresented while others may remain untouched. It is generally easier to prepare objective items on some topics than on others. And it is easier to prepare items that require the recall of simple facts than to devise items calling for critical evaluation, integration of different facts, or application of principles to new situations. Yet follow-up studies have shown that the factual details learned in a course are most likely to be forgotten, while the understanding of principles and their application to new situations show either no retention loss or an actual gain with time after completion of a course. Thus, the test constructed without a blueprint is likely to be overloaded with relatively impermanent and less important material. Many of the popular criticisms of objective tests stem from the common overemphasis of rote memory and trivial details in poorly constructed tests.

To guard against these fortuitous imbalances and disproportions of item coverage, *test specifications* should be drawn up before any items are prepared. For classroom examinations, such specifications should begin with an outline of the objectives of the course as well as of the subject matter to be covered. In listing objectives, the test constructor should ask himself what changes in behavior the course was designed to produce. Such changes may pertain to work habits, problem-solving styles, attitudes, interests, and other noncognitive characteristics, as well as to the acquisition of knowledge and the development of intellectual skills.

Table 30 illustrates a two-way specification table for a 50-item examination in economics, covering the topic of money and banking (Gronlund, 1970). In the left-hand column are listed four content areas or subject-matter categories to be covered. Across the top are four instructional objectives or types of learning to be tested. The entries in the body of the table give the desired number of items, classified by content and objective. These numbers should reflect the relative importance and scope of each topic and objective. When a particular objective or type of learning is inapplicable to a given topic, no items are listed in

TABLE 30
Two-Way Specification Table for a 50-Item Economics Test on Money and Banking

(Adapted from Gronlund, 1970, p. 42. © Copyright 1970 by Norman E. Gronlund. Reproduced by permission of Macmillan Publishing Co., Inc.)

Content Areas	Instructional Objectives				Total
	Knows Basic Terms	Understands Concepts and Principles	Applies Principles	Interprets Data	
Forms and functions of money	3	4	3		10
Operation of banks	4	3	5	3	15
Role of the Federal Reserve System	4	6	3	2	15
State regulation of banks	4	2	4		10
Total	15	15	15	5	50

the corresponding cell. Column and row totals show the overall weight of each topic and type of learning in the entire examination. Frequently, only these column and row totals are established in advance, thereby controlling the relative coverage of topics and types of learning, but not specifying the cross-classification of individual items.

The test constructor must also decide on the most appropriate *item form* for his material. The advantages traditionally cited in favor of essay questions are that they test the individual's ability to select, relate, and organize material, as well as his ability to express ideas clearly and accurately. Unfortunately, the time available to the examinee in answering essay questions is usually too short to measure these particular skills. It is possible, in fact, that essay examinations may be partly responsible for the habits of unclear and careless writing developed by many students. Apart from lack of time for effective organization and good writing, the student answering essay questions writes for an instructor who knows more about the subject than he does. Confident in the expectation that even obscure answers will be understood and duly credited, the student rarely takes the trouble to communicate clearly.

With small groups, it may not be worthwhile to prepare objective items, especially in an area that is rapidly changing and would require frequent item revision. Under such circumstances, essay questions that are carefully formulated and scored may provide the most practicable solution. Or a combination of essay and objective items may be chosen. Essay items can be improved by full and explicit formulation of the question and by systematic scoring procedures. A common weakness of essay examinations arises from subjectivity of scoring. It has been repeatedly demonstrated that the same answer may receive very different grades from different examiners or even from the same examiner at different times. To minimize these sources of error variance, it is advisable to list in advance the points to be covered by the answer and the credit to be assigned to each. Preparing sample answers also helps, especially when several persons are to do the scoring.

As the art of item writing develops, more and more of the abilities formerly believed to be amenable only to essay questions are proving to be testable by objective items. Among the chief advantages of objective items are ease, rapidity, and objectivity of scoring. Since each objective item requires much less of the examinee's time than does a typical essay question, objective items also permit a fuller coverage of content and hence reduce an important source of chance errors in total scores.

Among objective items, there is a choice of several specific forms, such as true-false, multiple-choice, completion, matching, and arrangement (in order of magnitude, chronological order, etc.). The content and type of learning to be tested would largely determine the most appropriate item form. Multiple-choice items have proved to be the most widely applicable. They are also easier to score than certain other forms, and reduce the chances of correct guessing by presenting several alternative responses.

Many practical rules for effective *item writing* have been formulated on the basis of years of experience in preparing items and empirical evaluation of responses. Anyone planning to prepare objective items would do well to consult one of the books summarizing these suggestions (e.g., Adkins, 1974; Ebel, 1972a, Chs. 5–8; Gronlund, 1976, Chs. 6–11; Nunnally, 1972, Chs. 5–7; Stanley & Hopkins, 1972, Chs. 8–10; Wesman, 1971). The *Taxonomy of Educational Objectives* (Bloom et al., 1956; Krathwohl et al., 1964), cited in Chapter 6 in connection with content validity, contains many examples of items designed to test different learning skills. Such published achievement tests as the Sequential Tests of Educational Progress (STEP) and the Stanford Achievement Test are also a good source of novel and imaginative item types.

To add one more summary of item-writing "rules" to the many already available would be redundant. However, a few examples will be given to suggest the kind of pitfalls that await the unwary item writer. Ambiguous or unclear items are a familiar difficulty. Misunderstandings are likely to occur because of the necessary brevity. It is difficult to write a single sentence that can stand alone with clarity and precision. In ordinary writing, any obscurity in one sentence can be cleared away by the sentences that follow; but it requires unusual skill to compose isolated sentences that can carry the whole burden unaided. The best test of clarity under these circumstances is to have the statement read by someone else. The writer, who knows the context within which he framed the item, may find it difficult to perceive other meanings in it. If he reads it at a later time, however, the writer himself may be able to spot ambiguities.

While students who feel they may have been cheated out of one or two score points by ambiguous items are quick to call it to the instructor's attention, the opposite type of error is less likely to be publicized. Yet many poorly constructed items give an advantage to the observant guesser. For example, one item may give away the answer to another occurring in a different part of the examination. Thus, one item may require the student to identify the name of the psychologist who developed the test of primary mental abilities, while another begins with the words, "In Thurstone's tests of primary mental abilities. . . ." Grammatical cues, such as differences in number or tense of verbs or the use of the indefinite article "a" instead of "an," may reveal the correct alternative in a multiple-choice item or may at least permit the student to eliminate one of the wrong alternatives.

In the effort to be clear, the inexperienced item writer frequently makes the correct alternative longer or qualifies it more fully than the other alternatives. Examinees soon learn that the odds are in their favor if they choose such an alternative when they do not know the correct answer. A final example is provided by clang, or alliterative associations, as in the following vocabulary item:

(poor) illicit: secret illegal sexual ignorant daring
(good) illicit: secret unlawful sexual illiterate daring

In the first example, the individual who does not know the meaning of *illicit* could get the item right by choosing the word that sounds most nearly like it. Substituting *unlawful* for *illegal* in the second version has eliminated this cue. In addition, the alliterative cue in this version would mislead the guesser, since on this basis he would choose a wrong alternative, *illiterate.*

DIAGNOSTIC AND CRITERION-REFERENCED TESTS

In contrast to the survey tests represented by both general achievement batteries and achievement tests in special areas, diagnostic tests are designed to analyze the individual's specific strengths and weaknesses in a subject and to suggest causes of his difficulties. Diagnostic tests have much in common with criterion-referenced tests. Although the latter are used for a variety of purposes, as described in Chapter 4, diagnosis is certainly one of their major functions. Like traditional diagnostic tests, criterion-referenced tests elicit information about the individual's performance in highly specific skills and relate this information to instructional prescriptions. Nearly all published tests in both the diagnostic and the criterion-referenced categories deal with either reading or mathematics.

DIAGNOSTIC TESTS. Available instruments for the diagnosis of *reading* difficulties vary widely in the thoroughness of analysis they permit and in the specific procedures followed. They range from group tests yielding two or three subtest scores, which serve little more than a survey function, to intensive clinical programs for individual case studies. Some provide detailed checklists of specific types of errors. The individual batteries may employ apparatus, such as tachistoscopes for controlling rate of exposure of printed material, and techniques for photographing the individual's eye movements while he reads.

Among the most common weaknesses of diagnostic reading tests are inadequate reliabilities coupled with high intercorrelations of the subtests from which separate scores are derived. Especially in some of the shorter group tests, these two conditions reduce the diagnostic or differential effectiveness of the tests. Several tests have also been criticized because of the superficiality of understanding required by the reading comprehension subtests. The measurement of rate of reading likewise presents special problems. Rate of reading depends on such factors as the difficulty of the material and the purpose for which it is being read. Any individual may thus have, not one, but many reading rates. This is more likely to be true of the experienced reader, who adjusts his rate to the nature of the task. Furthermore, in any one reading test, the response set established by the directions may differ widely from one person to another. For example, when told to read a passage carefully so as to be able to answer questions about it later, some persons may skim the material rapidly, while others will try to understand and memorize specific details.

The Stanford Diagnostic Reading Test is an example of a well-constructed group test. Spanning grades 2.5 to 13, this test is available in three levels. Several scores are obtained at each level. For instance, Level II, applicable from the middle of grade 4 to the middle of grade 8, comprises six tests which yield eight scores as follows:

1. Reading Comprehension
    - Literal
    - Inferential
    - Total
2. Vocabulary (auditory)
3. Syllabication
4. Sound Discrimination
5. Blending
6. Rate of Reading

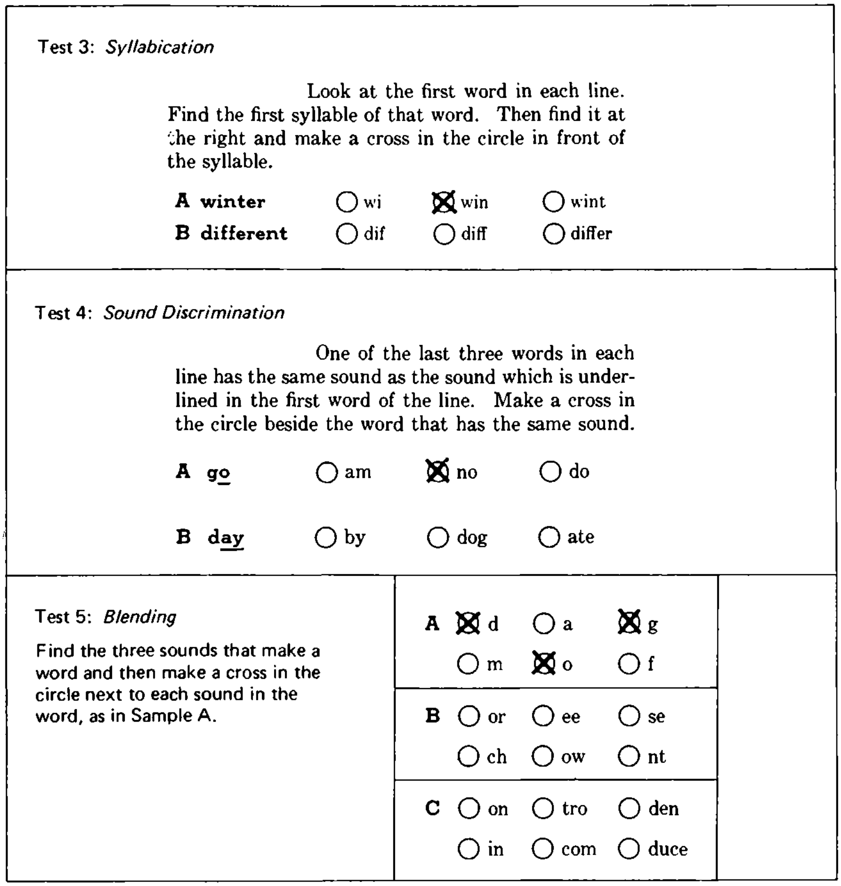
Sample items from Tests 3, 4, and 5, covering word recognition skills, are shown in Figure 72.

Scores on each subtest of the Stanford Diagnostic Reading Test may be interpreted in terms of within-grade stanines or percentiles, although the stanine conversion is recommended and is employed in the pupil profile charts. Reading comprehension scores may also be translated into grade scores[3] to provide an index of the pupil's overall reading performance. The manual and the Instructional Placement Reports provide supplementary discussion to aid in the qualitative interpretation of performance patterns, together with appropriate instructional suggestions. In the standardization program, the Stanford Diagnostic Reading Test was administered to approximately 12,000 children in six school systems, who also took the reading subtests of the Stanford Achievement Test. The normative sample was selected in terms of the latter scores so as to be representative of the standardization sample of the Stanford Achievement Test.

Subtest reliabilities of the Stanford Diagnostic Reading Test are generally satisfactory. Split-half reliabilities within single-grade groups cluster in the high .80's and .90's. Alternate-form reliability of the rate-of-reading subtest at Level II was .72. Intercorrelations of the subtests vary widely with the particular tests correlated and tend to be lower at the higher grades. For Level II, the median correlations are in the .40's and .50's. When considered together with the reliability coefficients, these in-

---

[3] Grade scores are simply grade equivalents without the decimal point. For example, a grade score of 54 corresponds to a grade equivalent of 5.4, representing the fourth month of grade 5.



Test 3: *Syllabication*

Look at the first word in each line. Find the first syllable of that word. Then find it at the right and make a cross in the circle in front of the syllable.

A winter ○ wi ☒ win ○ wint

B different ○ dif ○ diff ○ differ

Test 4: *Sound Discrimination*

One of the last three words in each line has the same sound as the sound which is underlined in the first word of the line. Make a cross in the circle beside the word that has the same sound.

A go ○ am ☒ no ○ do

B day ○ by ○ dog ○ ate

Test 5: *Blending*

Find the three sounds that make a word and then make a cross in the circle next to each sound in the word, as in Sample A.

A ☒ d ○ a ☒ g
○ m ☒ o ○ f

B ○ or ○ ee ○ se
○ ch ○ ow ○ nt

C ○ on ○ tro ○ den
○ in ○ com ○ duce

FIG. 72. Sample Items from Three Subtests of the Stanford Diagnostic Reading Test, Level II. (A revised edition is scheduled for 1977.)

(Copyright © 1966–69, 1973, by Harcourt Brace Jovanovich, Inc. Reproduced by special permission of the publisher.)

tercorrelations indicate that the subtests should differentiate reasonably well among the individual's reading skills.

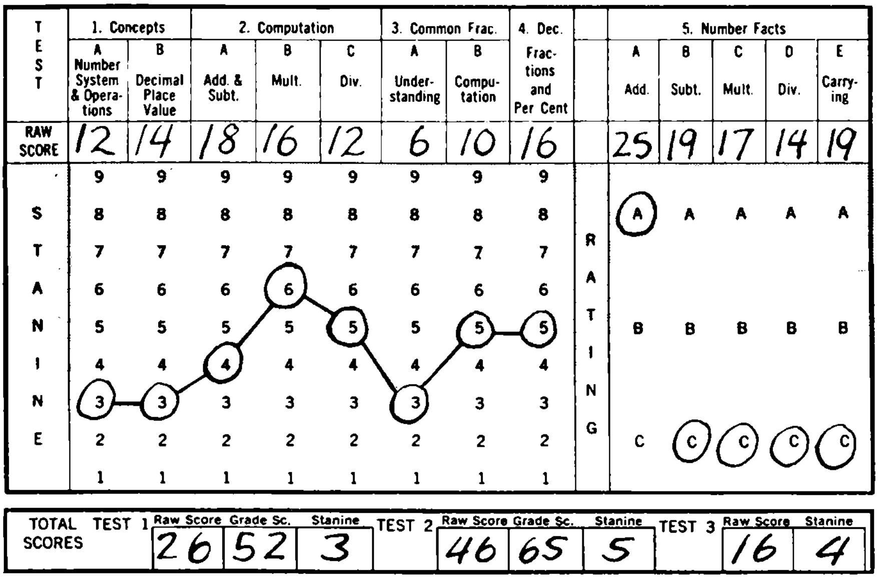
A typical battery for intensive individual testing is the Durrell Analysis of Reading Difficulty, designed for grades 1 to 6. The Durrell tests utilize series of paragraphs graded in difficulty, a set of cards, and a simple tachistoscope. Scores are provided for rate and comprehension of oral and silent reading, listening comprehension, rapid word recognition, and word analysis. Supplementary tests of written spelling and speed of handwriting are included. For nonreaders, there are measures of visual memory for word forms, auditory analysis of word elements, letter recognition, rate of learning words, and listening comprehension. All of the comprehension tests actually require simple recall of details read or heard and do not call for much understanding. One of the chief contributions of this battery is a checklist based on reading errors identified in a survey of 4,000 children. The battery is better suited for qualitative than for quantitative analysis of performance. Suggestions for remedial teaching are also provided in the manual.

The diagnostic testing of *mathematical skills* presents several of the same methodological problems encountered in diagnostic reading tests. The testing procedures also parallel those in the reading area, ranging from fairly short group tests to intensive programs for individual examination. The number of diagnostic tests developed in the mathematics area, however, is much smaller; and formally organized remedial programs in this area are fewer than in reading.

The Stanford Diagnostic Arithmetic Test closely parallels the previously discussed Stanford Diagnostic Reading Test in both administration and test-construction procedures. Available in two levels covering grades 2.5 to 8.5, it utilizes group testing to provide a profile of performance in specific arithmetic skills. The profile of scores obtained by a sixth-grade pupil is reproduced in Figure 73. In it can be seen the names of the subtests and the types of normative conversions applied to each score. The stanine is the principal type of score employed, with grade scores also reported for total scores in arithmetic concepts and computation. Since distributions of scores on the last subtest, Number Facts, are highly skewed, only a broad classification of scores into three letter grades is given. The Number Facts subtest, which requires rapid computations with orally presented numbers, was designed to assess the pupil's ability to recall simple number facts, such as $4 \times 7 = 28$. The same number combinations are used in the corresponding items in addition and subtraction, and in multiplication and division. For instance, if item 1 in addition is $8 + 2 = 10$, item 1 in subtraction is $10 - 2 = 8$. Hence, it is possible to see whether or not a pupil recalls related number facts.

The recommended testing schedule for Level II includes seven sittings, varying in length from 15 to 53 minutes. Several subtests are given orally by the teacher; the entire test requires little reading, especially at the lower levels. Technical properties of the test, with regard to norms, reliability, and intercorrelations of subtests, are quite similar to those of the Stanford Diagnostic Reading Test.

To illustrate a very different approach, we may consider an early classic that has enjoyed wide popularity. The Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic requires individual administration, since each problem is solved orally by the pupil. It is thus possible to observe the work methods employed by the child in carrying out the different operations. Errors as well as undesirable work habits



TEST	1. Concepts		2. Computation			3. Common Frac.		4. Dec.		5. Number Facts				
	A Number System & Operations	B Decimal Place Value	A Add. & Subt.	B Mult.	C Div.	A Understanding	B Computation	Fractions and Per Cent		A Add.	B Subt.	C Mult.	D Div.	E Carrying
RAW SCORE	12	14	18	16	12	6	10	16		25	19	17	14	19
	9	9	9	9	9	9	9	9						
S	8	8	8	8	8	8	8	8		[✓ A]	A	A	A	A
T	7	7	7	7	7	7	7	7	R					
A	6	6	6	[✓ 6]	6	6	6	6	A					
N	5	5	5	5	[✓ 5]	5	[✓ 5]	[✓ 5]	T	B	B	B	B	B
I	4	4	[✓ 4]	4	4	4	4	4	I					
N	[✓ 3]	[✓ 3]	3	3	3	[✓ 3]	3	3	N					
E	2	2	2	2	2	2	2	2	G	C	[✓ C]	[✓ C]	[✓ C]	[✓ C]
	1	1	1	1	1	1	1	1						

TOTAL SCORES	TEST 1 Raw Score	Grade Sc.	Stanine	TEST 2 Raw Score	Grade Sc.	Stanine	TEST 3 Raw Score	Stanine
	26	52	3	46	65	5	16	4

FIG. 73. Pupil Profile of a Sixth-Grade Girl on Stanford Diagnostic Arithmetic Test, Level II. This pupil learns well what she is taught about arithmetic computation but has difficulty in understanding the basic concepts of the subject. She has also failed to master the number facts but is able to figure them out if given time as on Test 2. (A revised edition is scheduled for 1977.) (Reproduced by permission of Harcourt Brace Jovanovich, Inc.)

are recorded on a checklist, which includes the most common difficulties encountered in arithmetic. The problems were specially chosen so as to elicit such difficulties when present. This test is applicable in grades 2 to 8. There are no time limits, no norms, and no total score, the problems and checklist being designed for qualitative rather than quantitative analysis of arithmetic performance. It should be added that the development of criterion-referenced tests, especially when combined with computer utilization, is opening the way for sophisticated techniques of error analysis in group-administered arithmetic tests (see, e.g., Glaser & Nitko, 1971, p. 645).

In connection with the use of all diagnostic tests, one point deserves special emphasis. The diagnosis of reading and arithmetic difficulties and the subsequent program of remedial teaching are the proper functions of a trained specialist. No battery of diagnostic tests could suffice for this purpose. The diagnosis and treatment of severe reading disabilities require an intensive clinical case study, including supplementary information on sensory capacities and motor development, medical and health history, complete educational history, data on home and family background, and a thorough investigation of possible emotional difficulties. In some cases, serious reading retardation proves to be only one symptom of a more basic personality maladjustment. Although survey and group diagnostic tests may serve to identify individuals in need of further attention, the diagnosis and therapy of reading disabilities often represent a problem for the clinician. This may also be true in cases of severe arithmetic disabilities. The appropriate procedures will be considered further in Chapter 16, in connection with clinical testing.

CRITERION-REFERENCED TESTS. Several published criterion-referenced tests are now available for reading and arithmetic and, thus far, one for study skills (see Appendix E for list). At this stage, all are designed for the elementary school grades. Items are described in terms of highly specific objectives, the number of such objectives varying from 30 to nearly 200 in any one test. Mastery scoring is used regularly, the individual's performance being judged as indicating mastery or nonmastery of each skill. The tests are diagnostic and prescriptive, the performance records being keyed to basic texts, workbooks, and supplementary instructional materials to which the student is directed. Usually the instruments are organized into a continuing testing program, so that pupils can be examined after learning each skill.

A common pattern consists of a preliminary broad-range test, which may contain as few as one or two items for each objective, and several follow-up tests for the more intensive testing of a single objective or a group of closely related objectives. An example is the combination of Prescriptive Reading Inventory and Interim Evaluation Tests published by CTB/McGraw-Hill. Another is the Skills Monitoring System developed by Harcourt Brace Jovanovich. At each grade level, this system includes a skill locator and 40 to 60 one-page minitests consisting of 8 to 12 items on a single objective. An interesting feature of this particular system is that the minitests are self-scored and provide immediate feedback to the pupils. This is accomplished by the use of chemically treated crayons and latent images. When the examinee fills in the box next to the answer he chooses, "Yes" appears if his choice is correct; if not, he continues to mark other alternatives until he finds the correct answer.

Most test publishers have been grappling with the technical problems characteristic of criterion-referenced tests, with varying degrees of success. There is the three-headed problem of assessing a long list of narrowly defined objectives, including enough items to measure each objective reliably, and keeping the whole process within manageable length. There is the perennial problem of setting an acceptable performance level as a standard of mastery. And there are the special problems associated with the measurement of reliability and validity of such instruments.

It is unlikely that a single set of highly specific instructional objectives would suit all users, except possibly at the most elementary stages of learning. It would seem, too, that one of the intrinsic advantages of criterion-referenced testing stems from its being tailored to local needs. For this reason, some test publishers, including Educational Testing Service and Houghton Mifflin, have established services for the construction of custom-made criterion-referenced tests to fit local educational needs. These services include the collaboration of measurement and curriculum experts with local staff in the development and installation of the testing program; and the availability of item banks comprising many thousands of edited, reviewed, and pretested questions on a wide range of academic subjects.

SPECIALIZED PROGNOSTIC TESTS

Certain types of tests designed for use in educational contexts are essentially predictive instruments. As such, they function as aptitude rather than as achievement tests. At the same time, they frequently resemble achievement tests closely in content, because what they undertake to predict is usually performance in a specific academic course. The extent to which the individual has acquired prerequisite knowledge and skills is a good predictor of his subsequent performance in such a course. How well he learns new material similar to that which he will encounter in the course is another.

Prognostic tests in mathematics are illustrated by the Orleans-Hanna Algebra Prognosis Test and the Orleans-Hanna Geometry Prognosis Test. In both of these tests, the student is provided with simple material to learn from algebra or geometry, and is immediately tested on what he has learned. These tests are thus worksamples, in which the student's subsequent course learning is predicted from his performance in the sample learning tasks. Other prognostic tests in mathematics cover a combination of prerequisite arithmetic skills and new learning. Some contain material similar to that found in the numerical subtests of intelligence tests, such as number series completions. All such prognostic tests are normally validated against subsequent course grades and terminal achievement test scores.

Another type of predictive instrument is illustrated by the Modern Language Aptitude Test. Designed to assess the capacity of an English-speaking student for learning any foreign language, this test utilizes both paper-and-pencil and tape-recorded materials. It is suitable for high school, college, and adult groups. Two of its subtests require the learning of orally presented numbers and visually presented words in an artificial language. The other three parts test the person's sensitivity to English grammatical structure, as well as certain word recognition skills with visual and auditory materials. The test may also be administered in a shorter form not requiring a tape recorder.

Percentile norms are reported for beginning language students in grades 9, 10, 11, and 13, as well as for military and civilian personnel assigned to intensive foreign language training. Most of the normative samples are small, however, and may not be comparable at different grade levels. Split-half reliabilities for both total test and short form are high, most of them exceeding .90. Data on predictive validity in college and high school groups appear promising. In an earlier experimental version, this test proved especially effective in predicting success in intensive language training courses conducted by the Foreign Service Institute of the Department of State, the Air Force, and the Army Language School. An elementary form of this test, suitable for grades 3 to 6, is also available.

Designed for the same general purpose, the Pimsleur Language Aptitude Test is applicable in grades 6 to 12. The total score on this test is a weighted sum of the scores on four subtests, the student's grade-point average in academic areas other than foreign languages, and his self-rated interest in learning a foreign language. The four subtests provide measures of English vocabulary, ability to identify and correctly apply grammatical endings and word sequences in an artificial language (see Fig. 74), ability to learn new phonetic distinctions and to recognize them in

The list below contains words from a foreign language and the English equivalents of these words.

gade	father, a father
shi	horse, a horse
gade shir le	Father sees a horse.

By referring to the above list, figure out how the following statement should be expressed in this language.

A horse sees Father.

(Correct answer: shi gader le.)

FIG. 74. Demonstration Item from Pimsleur Language Aptitude Test.
(Reproduced by permission of Harcourt Brace Jovanovich, Inc.)

different contexts, and ability to associate sounds with their written symbols. The directions for all tests as well as the items for the last two tests are on tape.

Both percentile and stanine norms are given for students tested at the beginning of Grades 7, 8, and 9, as well as for students completing the first year of instruction in French or Spanish. Each normative sample includes from about 1,000 to about 3,000 students drawn from 5 to 13 states. Norms are available for total scores and for verbal and auditory scores, each of the latter being based on the sum of two subtests. Split-half reliabilities in the .80's were obtained for total scores within each grade level of the normative sample. A few medium to high correlations with terminal course grades and achievement test scores are reported. It is likely, however, that the parts of the battery should be combined with different weights for predicting different aspects of language mastery, such as listening comprehension, speaking, reading, and writing.

ASSESSMENT IN EARLY CHILDHOOD EDUCATION

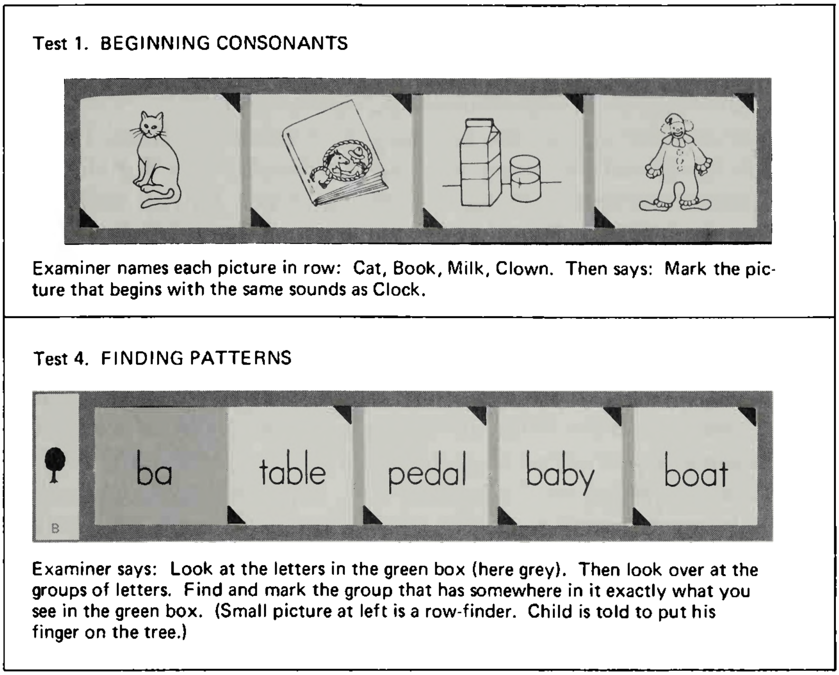
Many new instruments for measuring the educational development of young children have been published since 1970. Several influences have contributed both to the amount and nature of this activity. Research in early cognitive development, the burgeoning of programs for preschool education, and the widespread concern about the effects of cultural handicaps on the child's ability to profit from school instruction have all played major roles. The Piagetian approach and criterion-referenced orientations can be seen in several recently developed instruments. Some of the tests were designed principally to measure the outcomes of early childhood education and thus function as achievement tests. Others are presented as predictive or aptitude instruments to assess the child's readiness for first-grade instruction. The two types of instruments merge imperceptibly, however, and each can usually serve either purpose.

School readiness refers essentially to the attainment of prerequisite skills, knowledge, attitudes, motivations, and other appropriate behavioral traits that enable the learner to profit maximally from school instruction. These prerequisites are what Hunt and Kirk (1974) have called the "entry skills" that the child needs to cope with the teaching-learning situation encountered in the first grade. At one time, readiness was conceived largely in terms of maturation. To be sure, the development of certain minimum physical qualifications facilitate some kinds of learning. Unless a child can make the necessary auditory discriminations, he cannot learn to speak by the usual procedures; without the ability for fine motor coordination, he is unable to manipulate a pencil in writing. Most school learning, however, is not so closely linked to sensorimotor development. In the mastery of educational tasks, the importance of prior learning is being increasingly recognized. More and more emphasis is now placed on the hierarchical development of knowledges and skills, whereby the acquisition of simple concepts equips the child for the learning of more complex concepts at any age.

Readiness tests are generally administered upon school entrance. While they have much in common with intelligence tests for the primary grades, readiness tests place more emphasis on the abilities found to be important in learning to read. Some attention is also given to the prerequisites of numerical thinking and to the sensorimotor control required in learning to write. Among the specific functions frequently covered are visual and auditory discrimination, motor control, aural comprehension, vocabulary, quantitative concepts, and general information.

A widely used readiness battery is the Metropolitan Readiness Tests (MRT). In its 1976 edition, this battery has two levels, one for the beginning or middle of kindergarten and one for the end of kindergarten or the beginning of Grade 1. Both levels are orally administered, requiring the child to make simple marks in the test booklet. A Practice Booklet is administered in advance to acquaint the child with the materials and procedures he will encounter in the tests themselves. Level II contains the following eight tests, of which the last two are optional:

1. *Beginning Consonants:* child looks at four pictures in test booklet, while examiner names each; then he selects the one whose name begins with the same sound as a word pronounced by examiner but not pictured (Fig 75).
2. *Sound-Letter Correspondence:* each item consists of a picture followed by four letters; after examiner names the picture, child selects the letter that stands for the sound at the beginning of the name.
3. *Visual Matching:* match what is shown at the beginning of each row with one of four choices which follow; includes letter series (nonwords), words, numerals, and letter-like forms (artificial letters).
4. *Finding Patterns:* designed to assess ability to "see" a given formation of letter groups, words, numerals, or artificial letters within a larger grouping; child must perceive the visual stimulus, then match it with its hidden counterpart amid distracting surroundings (Fig. 75).
5. *School Language:* tests child's understanding of some of the basic and some of the more complex grammatical structures and concepts of the usual language of school instruction. E.g., "Mark the picture that shows this: The car that is in front of the truck is turning the corner."
6. *Listening:* tests comprehension of vocabulary and structure of orally presented passages; some items call for inferences and conclusions.



Test 1. BEGINNING CONSONANTS

Examiner names each picture in row: Cat, Book, Milk, Clown. Then says: Mark the picture that begins with the same sounds as Clock.

Test 4. FINDING PATTERNS

Examiner says: Look at the letters in the green box (here grey). Then look over at the groups of letters. Find and mark the group that has somewhere in it exactly what you see in the green box. (Small picture at left is a row-finder. Child is told to put his finger on the tree.)

FIG. 75. Demonstration Items from the Metropolitan Readiness Tests. Throughout the battery, the child is instructed to mark his answer by drawing a line from one black corner to the other, a procedure he was taught to follow in the practice booklet and which is demonstrated anew on the chalkboard before the administration of each test. This marking technique permits computer scoring of the booklets.

(Copyright © 1974. Items reprinted by courtesy of Harcourt Brace Jovanovich, Inc.)

7. *Quantitative Concepts* (optional): tests knowledge of concepts basic to an understanding of primary level mathematics, including size, shape, position, quantity, etc.
8. *Quantitative Operations* (optional): assessment of child's mastery of counting and simple mathematical operations.

Scores on either level of the MRT can be converted to stanines. The norms were derived from a representative national sample of about 100,000 kindergarten and first-grade children. Both split-half and alternate-form reliabilities are high. As is true of readiness tests generally, this test shows considerable overlap with intelligence tests for the primary grades. Several studies with the earlier edition of the MRT yielded correlations in the .70's with group intelligence tests. However, when checked against end-of-year achievement test scores, readiness tests yield consistently higher validities than do intelligence tests.[4] Since readiness tests are specifically designed to assess identifiable prerequisite skills, it is understandable that they are better predictors of school achievement than are the more generally oriented intelligence tests.

A different approach to the assessment of school readiness is represented by the Gesell Developmental Tests, assembled by Ilg and Ames at the Gesell Institute (Ilg & Ames, 1965; Ilg, Ames, & Apell, 1965). The procedure described by these authors is essentially that of a clinical examination requiring a trained examiner. Ilg and Ames argue that the usual intelligence or school readiness test is too limited in coverage and that a child may obtain a superior score on such a test but be immature in his general behavior. In contrast, their objective is to assess the developmental level at which the child is behaving as a total organism. Lasting approximately a half hour, the examination begins and ends with a brief interview and includes predominantly perceptual-motor tests and a few tests requiring oral responses. As adjuncts to evaluation, the authors also recommend an examination of the teeth, with special reference to the loss of baby teeth as an index of physical development; a vision examination; oral reading tests; and a projective test.

Detailed instructions are provided for administering the tests and evaluating the responses. Norms are in the form of percentages of children at each age giving each response; but no quantitative score is computed. A further interpretive aid consists of capsule portraits, including a general description of the behavior characteristic of each age and a summary of typical test responses. The normative samples were small and drawn from a narrowly limited population. Although not attempting to evaluate these basically clinical procedures in psychometric terms, Ilg and Ames cite some promising evidence of predictive validity, especially at the kindergarten and first-grade levels. In a later, independent study, a modified version of the Gesell Developmental Tests administered to 80 kindergarten children yielded a validity coefficient of .64 with total score on the Stanford Achievement Test obtained at the end of the first grade (Kaufman & Kaufman, 1972). This predictive validity was higher than that found for a standardized group intelligence test administered in the same study.

The current emphasis on direct assessment of early childhood education is reflected in the Stanford Early School Achievement Test (SESAT). The manual states that "SESAT is not intended to be a 'readiness test,' except in the same sense that an achievement test at Grade 1 is a readiness test for Grade 2 . . . . SESAT results are to be built upon, not waited upon" (SESAT Level I Manual, p. 12). Very true, but we must not be misled by superficial differences in nomenclature. The Metropolitan Readiness Tests, for example, correlate very highly with SESAT and may be used effectively for the same purposes as SESAT.

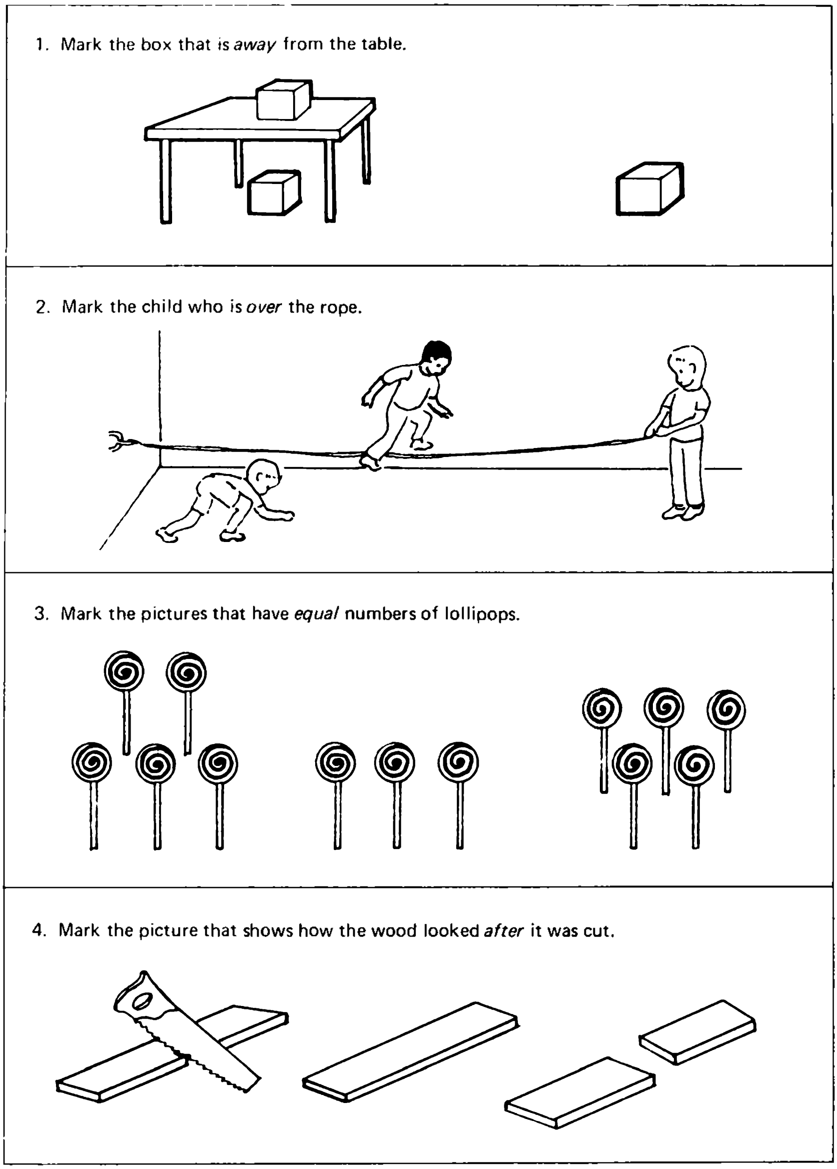
---

[4] Unpublished data covering both MRT and other readiness tests. Assembled by Blythe C. Mitchell, Harcourt Brace Jovanovich (Personal communication, July 1974.)

It will be recalled that SESAT represents a downward extension of the Stanford Achievement Test and is tied into the same score scale. Level I, suitable for kindergarten and the beginning of Grade 1, includes four separately scored tests: (1) The Environment (information about social studies and science), (2) Mathematics, (3) Letters and Sounds, and (4) Aural Comprehension. Level II, designed for Grade 1, comprises six tests, four corresponding to those in Level I and two additional tests, Word Reading and Sentence Reading. The manuals contain suggestions for instruction in any area in which the tests reveal deficiencies. Puerto Rican Spanish and Chicano Spanish editions have been prepared, with translated oral instructions. The manuals warn, however, that comparability of these adapted versions with the standard version cannot be assumed. They should therefore be used only with local norms.

Some tests have combined a criterion-referenced approach with norm-referenced interpretations. Examples include the Tests of Basic Experiences (TOBE) and the Boehm Test of Basic Concepts. Of the two, TOBE has a broader scope, covering a rather heterogeneous collection of items grouped into four tests at each level. The Boehm test, on the other hand, was designed to serve a more narrowly circumscribed function. According to the author, its purpose is "to assess beginning school children's knowledge of frequently used basic concepts widely but sometimes mistakenly assumed to be familiar to children at their time of entry into kindergarten or first grade" (Manual, p. 3). Each of the two equivalent forms of the Boehm test contains 50 items, classified under Space (location, direction, orientation, dimensions), Quantity (including number), Time, and Miscellaneous. Four typical items are shown in Figure 76.

A well-designed class record form facilitates the evaluation of test performance for both individuals and class in terms of total scores (norm-referenced) and individual items (criterion-referenced). A portion of this record form is reproduced in Figure 77. For each item, the form gives a key word identifying the concept and a letter indicating the broader classification. Each column shows the performance of the child whose name is written at the top. The child's total score and percentile are entered at the bottom of the column. Reading across each row permits the computation of the percentage of children in the group who answered each item correctly. For comparative purposes, the manual lists percentages of children in the standardization sample who passed each



1. Mark the box that is *away* from the table.

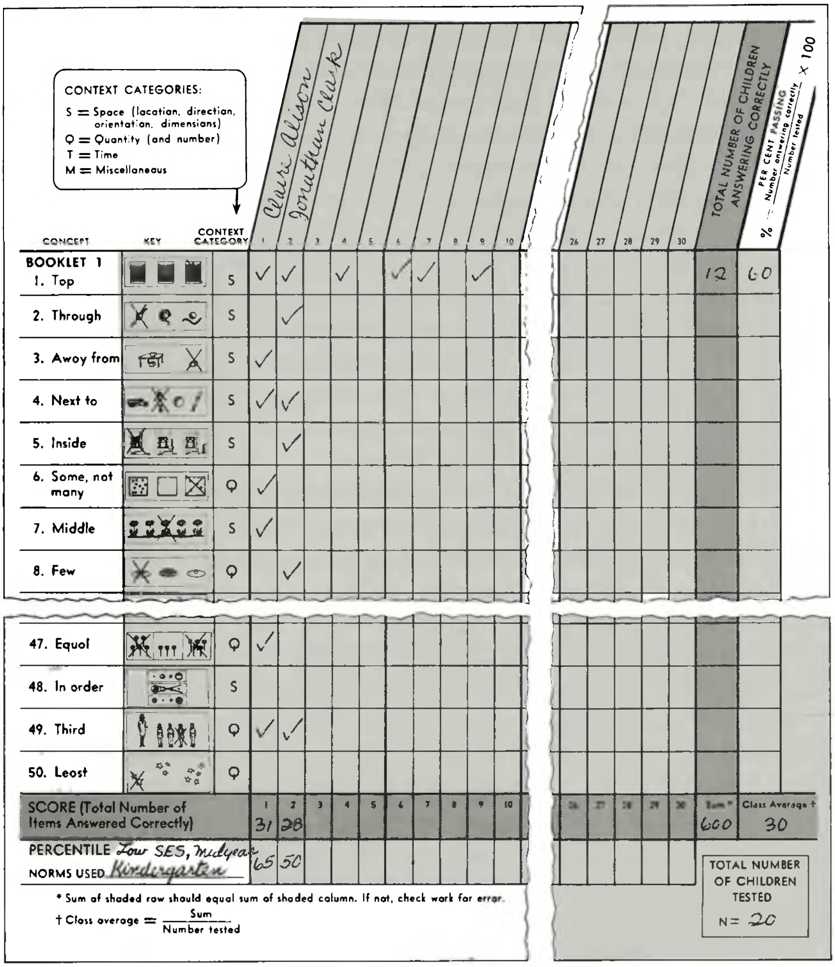
2. Mark the child who is *over* the rope.

3. Mark the pictures that have *equal* numbers of lollipops.

4. Mark the picture that shows how the wood looked *after* it was cut.

FIG. 76. Typical Items from Boehm Test of Basic Concepts.

(Reproduced by permission. Copyright © 1967, 1969, 1970 by The Psychological Corporation, New York, N.Y. All rights reserved.)



CONTEXT CATEGORIES:
S = Space (location, direction, orientation, dimensions)
Q = Quantity (and number)
T = Time
M = Miscellaneous

CONCEPT	KEY	CONTEXT CATEGORY	1 Claire Alison	2 Jonathan Clark	3	4	5	6	7	8	9	10	26	27	28	29	30	TOTAL NUMBER OF CHILDREN ANSWERING CORRECTLY	PER CENT PASSING % = Number answering correctly / Number tested × 100
BOOKLET 1 1. Top		S	✓	✓		✓		✓	✓		✓							12	60
2. Through		S		✓															
3. Away from		S	✓																
4. Next to		S	✓	✓															
5. Inside		S		✓															
6. Some, not many		Q	✓																
7. Middle		S	✓																
8. Few		Q		✓															
47. Equal		Q	✓																
48. In order		S																	
49. Third		Q	✓	✓															
50. Least		Q																	
SCORE (Total Number of Items Answered Correctly)			1 31	2 28	3	4	5	6	7	8	9	10	26	27	28	29	30	Sum* 600	Class Average† 30
PERCENTILE NORMS USED Low SES, midyear Kindergarten			65	50															

* Sum of shaded row should equal sum of shaded column. If not, check work for error.

† Class average = Sum / Number tested

TOTAL NUMBER OF CHILDREN TESTED
N = 20

FIG. 77. Part of Class Record Form for Boehm Test of Basic Concepts.

(Reproduced by permission. Copyright © 1967, 1969 by The Psychological Corporation, New York, N.Y. All rights reserved.)

item in high, middle, and low socioeconomic levels within kindergarten, Grade 1, and Grade 2 groups. Suggestions for individual remediation and group instruction are included in the manual.

Several of the concepts assessed in both the Boehm test and the TOBE are reminiscent of Piagetian scales. A more direct application of Piagetian methods is demonstrated in a series of individually administered tests described by Hunt and Kirk (1974). Although not a published instrument, this test series is fully described by the authors, who offer it as a paradigm to illustrate the applicability of Piagetian procedures to the determination of school readiness. The concepts tested pertain to color, position, shape, and number. Mastery of these concepts is tested through perceptual, listening, and spoken identification. For example, the child selects blocks that match a designated block in shape, in imitation of the examiner's demonstration; he responds to the examiner's naming of each shape by touching the correct block, giving it to the examiner, or carrying out other specified motor responses; and he himself names the shapes indicated by the examiner. A further test concerns the childs' ability to communicate these concepts orally to another child. Preliminary research with these tests revealed marked differences between the performance of culturally disadvantaged children in a Head Start Program and the performance of children in a middle-class nursery school. These findings support the hypothesis that children from different socioeconomic backgrounds vary substantially in their mastery of basic cognitive skills at the time of school entrance.

The Piagetian approach is also central to Let's Look at Children, a combined assessment and instructional program developed at Educational Testing Service. Originating in 1965 in a joint project of ETS and the New York City Board of Education dealing with first-grade children, the program was subsequently extended downward to prekindergarten and upward to Grade 3. Let's Look at Children is highly flexible, comprising an unstructured set of procedures and materials that can be used as a whole or as separate components. It was designed principally to aid teachers in the continuing, day-by-day processes of observing, understanding, assessing, and fostering the intellectual development of young children. The teacher functions as an integral part of the program.

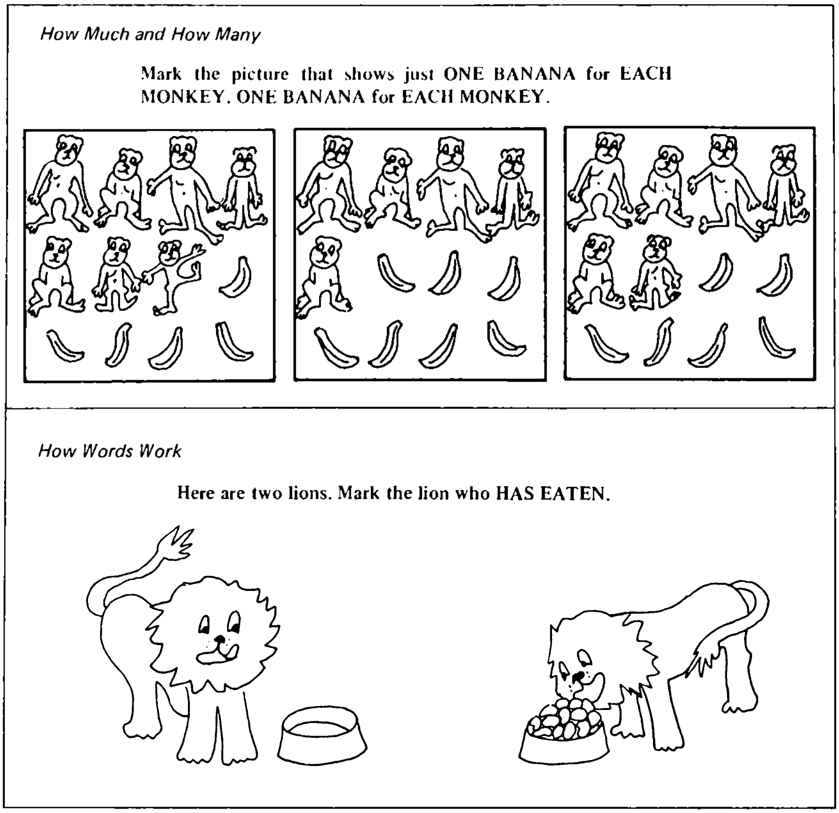
Essentially, this instrument consists of two volumes. Volume I, *Let's Look at Children: Theoretical Considerations,* introduces the reader to Piaget's formulation of conceptual development and discusses the nature and development of language and communication in the child. It provides a background for interpreting both children's classroom behavior and their responses to specific assessment procedures. Volume II, *Let's Look at Children: From Theory to Classroom,* outlines procedures for assessing many of the developmental processes described in the first volume. Usually these procedures also serve an instructional function insofar as they provide experiences relevant to the development of the corresponding skills and understandings. Most of the needed materials are available in the ordinary classroom; where special pictures or cards are required, they are supplied with Volume II. The procedures may be applied individually or with small groups of children. Guidelines for interpreting children's responses are included in Volume II, along with illustrative examples.

Several recent developments in psychometrics, psychological research, and early childhood education converge in Circus, a series of assessment instruments prepared by ETS for use with preschool and kindergarten children (*Circus,* 1973). The circus theme, which pervades the whole series, was chosen because of its intrinsic appeal to all children, regardless of sex, race, geography, or socioeconomic level. The entire series includes: 14 instruments for obtaining responses from children, only one of which requires oral responses and is therefore restricted to individual testing; two questionnaires for obtaining teacher ratings—one concerned with the child's habitual classroom behavior and the other with his behavior during the test situation; and an educational environment questionnaire, also filled out by the teacher, dealing with the child's familial and educational background, characteristics of the class and school environment, and the background, attitudes, and educational values of the teacher. In explaining the rationale for the last three instruments, the authors emphasize the importance of interpreting the child's cognitive development in relation to his own personality characteristics, as revealed in day-by-day behavior, and in the light of the environmental contexts in which he functions and develops (Messick, 1973a).

Below are a few examples of the 14 instruments requiring responses by the child:

1. *What Words Mean:* designed to assess the child's receptive vocabulary, the task is to choose the picture that corresponds to a given noun, verb, or modifier.
2. *How Much and How Many:* understanding of quantitative concepts. E.g., "Mark the cone with the most ice cream in it." (See also Fig 78)
6. *Noises:* child identifies picture that belongs with "real world" sounds from a tape cassette. E.g., baby's cry, crackling of a newspaper, dripping of a faucet, croaking of a frog.
8. *How Words Work:* discriminating between verb forms, prepositions, negatives and positives, sentence orders. (See Fig. 78)
10. *Say and Tell:* assesses productive as contrasted to receptive language skills, in three parts:
    (a) *Descriptive Usage.* E.g., child is handed two pennies and asked to tell all about them. "Suppose someone didn't know what they are. . . ."
    (b) *Linguistic Structure.* E.g., "Here is a house. Here are two ———."
    (c) *Narrative Usage.* Teacher shows child a large colored picture, explaining that it is a picture out of a storybook and adding, "I don't have the story that was in the book, so I want you to make up a story to go with the picture."

Users may choose any of the Circus instruments suited to their needs. All parts of the program are designed for use by classroom teachers. A preliminary practice device is employed to give children an opportunity



How Much and How Many

Mark the picture that shows just ONE BANANA for EACH MONKEY. ONE BANANA for EACH MONKEY.

How Words Work

Here are two lions. Mark the lion who HAS EATEN.

FIG. 78. Illustrative Items from Circus: An Assessment Program for Preprimary Children. Instructions are given orally by the examiner.

(Copyright © 1972, 1974 by Educational Testing Service. All rights reserved. Reprinted by permission.)

to learn the mechanics of marking responses and moving from item to item. Performance can be described in both norm-referenced and criterion-referenced terms for individuals as well as groups. Numerical scores are converted to "sentence reports," giving the child's strengths and weaknesses in specific functions. The accompanying manuals provide a theoretical framework for each measure to aid in the interpretation of results.

CHAPTER 15

Occupational Testing

PSYCHOLOGICAL tests are commonly employed as aids in occupational decisions, including both individual counseling decisions and institutional decisions concerning the selection and classification of personnel. Nearly every type of available test may be useful in occupational decisions. Multiple aptitude batteries and interest tests are particularly appropriate in counseling situations. Special aptitude tests have often been developed primarily for occupational purposes. Further discussion of the role of tests in counseling can be found in a number of sources. Super and Crites (1962) survey those tests that are most frequently employed by the counselor; Goldman (1971) provides an excellent introduction to the ways in which tests may be most effectively incorporated into the counseling process. An overview of the use of psychological tests in industry, with illustrations covering many specific occupations, is given by Lawshe and Balma (1966) and by McCormick and Tiffin (1974).

As in our treatment of educational tests in Chapter 14, we shall concentrate in this chapter on those tests specially designed for occupational purposes, over and above the more widely applicable instruments discussed in other chapters. We shall also examine the procedures typically followed in industrial settings to assemble and validate batteries of appropriate tests. And we shall consider the use of tests in professional training programs and in certification.

VALIDATION OF INDUSTRIAL TESTS

In the validity sections of test manuals, users are frequently cautioned that the validity data reported by the publisher should be rechecked and supplemented in local validation studies. This precaution is especially germane to the industrial use of tests, where any one test may be considered in connection with several hundred different jobs. It is unlikely that validity data pertaining to all these jobs were gathered by the test publisher. It is well known, too, that jobs bearing the same title, such as clerk or secretary, may entail vastly different job functions and responsibilities in different organizations. Moreover, the particular context in which the individual performs his functions often makes a difference. For example, the type of customer with whom a salesperson deals may substantially alter the optimum qualifications for the job.

From the standpoint of both employee and employer, it is obviously of prime importance that individuals be placed in jobs for which they have appropriate qualifications. Effective placement also implies that traits irrelevant to the requirements of the particular job should not affect selection decisions, either favorably or unfavorably. If a mechanical ability test requires a much higher level of reading comprehension than does the job, its use would not lead to the most effective utilization of personnel for that job. The simple psychometric fact that test validity must be ascertained for specific uses has long been familiar. It has now acquired new urgency because of the widespread concern about the job placement of culturally and educationally disadvantaged minorities (see Ch. 3 and Appendix B). An invalid test or one which includes elements not related to the job under consideration may unfairly exclude minority group members who could have performed the job satisfactorily.

COMPLETE LOCAL VALIDATION. Ideally, each test considered for personnel selection should undergo a full-scale local investigation of its criterion-related predictive validity. Specific validation procedures for this purpose were discussed in Chapters 6 and 7. In an industrial setting, such an investigation involves four principal steps:

(1) conducting a job analysis to identify the major job elements and specify the corresponding skills, knowledge, and other worker traits required by the job;

(2) selecting or constructing tests to assess the characteristics specified in step 1;

(3) correlating each test with appropriate criteria of job performance and choosing tests for the final battery;

(4) formulating a strategy for personnel decisions, i.e., determining how scores on the chosen tests will be used in making operational decisions.

The ideal procedure would further require the hiring of an unselected sample of applicants and their retention on the job for a period long enough to accumulate adequate and stable criterion data.

That such a full-scale, longitudinal validation study is unrealistic in the large majority of industrial situations is obvious. Even under unusually favorable conditions, with access to large employee samples, several practical limitations become apparent (see, e.g., Anastasi, 1972; Campbell, Crooks, Mahoney, & Rock, 1973). In Chapter 6, three principal reasons were given why such local investigations are likely to yield uninterpretable results: (1) unavailability of sufficiently dependable and inclusive criterion data; (2) inadequate number of employees performing the same or closely similar jobs; and (3) restriction of range through preselection, since not all applicants can be hired and followed up.

If time restrictions or other conditions necessitate the use of present employees rather than applicants, several additional features of the experimental design are sacrificed. Attrition over time will not only shrink sample size, but will also further reduce variability with a corresponding lowering of validity coefficients. Present employees will also have had varying amounts of job experience, which may be reflected in their test performance and thereby confound the relations between test score and criterion. Furthermore, when employees are given tests "for research purposes only," their motivation and test-taking attitudes may be quite unlike those of genuine job applicants; and these conditions are likely to be reflected in their test performance.

PROCEDURES UTILIZING JOB ANALYSIS AND EXTERNAL DATA. In view of the practical difficulties in the way of full-scale local investigations of criterion-related validity, a number of alternative procedures may be considered. These procedures can be ordered along a continuum representing degree of resemblance between test content and job functions. All depend upon a thorough and systematic job analysis. To be effective, a job analysis must be specific to the job. A description in terms of vague generalities that would be equally applicable to many jobs is of little use for this purpose. The job analysis must focus on those requirements that differentiate this job from other jobs. An effective job analysis should also concentrate on those aspects of performance that differentiate most sharply between the better and the poorer workers. In many jobs, workers of different levels of proficiency may differ little in the way they carry out most parts of their jobs—only certain features of the job may bring out the major differences between successes and failures. In his classic book on *Aptitude Testing*, Hull stressed the importance of these differentiating aspects of job performance, which he called "critical part-activities" (Hull, 1928, p. 286). Later, this concept was reemphasized by Flanagan (1949, 1954a), under the name of "critical requirements."

If it is to yield usable information, job analysis requires systematic data-gathering procedures. The job analyst should preferably combine several sources of information to obtain a well-rounded picture of the job activities. He may consult published sources, including descriptions of similar jobs as well as training and operating manuals prepared for the particular job. Performance records may be examined, particularly those containing qualitative descriptions of common errors, learning difficulties, and reasons for failure on this job. Records of customer complaints may provide clues useful in the analysis of sales jobs. An important source of information for any kind of job is provided by interviews—with supervisors, job instructors, and workers of varying degrees of experience and job success.

Other, more direct techniques of job analysis include activity logs kept by workers for a specified period of time, direct observation of a sample of workers by the job analyst, and (if practicable) performance of the job by the analyst himself. To implement the concept of critical requirements, Flanagan (1954a) proposed the critical incident technique. Essentially, this technique calls for factual descriptions of specific instances of job behavior that are characteristic of either satisfactory or unsatisfactory workers. A record of such incidents is usually kept by the supervisor during a designated period, such as two weeks. Growing out of Air Force research during World War II, the critical incident technique has been employed with such varied groups as commercial airline pilots, research personnel in physical science laboratories, factory workers, dentists, and department store salesclerks.

Job analysis enables the personnel psychologist to specify the critical worker requirements for success in a given job. The analysis may indicate, for example, that the job requires finger dexterity, perceptual speed, computational accuracy, and ability to work effectively under distracting conditions. With such detailed specifications of job elements, available tests designed to measure the appropriate functions may be utilized in finding *synthetic validity* against a composite criterion of job performance, as explained in Chapter 6.

If the job analysis reveals a few fairly broad critical abilities, it is likely that published tests can be located with adequately established *construct valdity* for the measurement of these aptitudes. Preferably there should be performance ratings or other criterion data on the separate job elements so that scores on each test can be correlated with the appropriate criterion data to obtain some local verification of the relation. In this approach, however, much depends on the accuracy and thoroughness of the job analysis and on the fullness with which the construct validity of the tests had previously been explored.

Some tests employ *simulation* to reproduce the functions performed on the job. Examples range from the operation of a miniature punch press to simulators for locomotive engineers and for airplane pilots. Simulators have been used for both training and testing purposes in the NASA space program, as well as in a number of military specialties. In a different context, simulation techniques have been applied to the design of tests for certain aspects of executive work (N. Frederiksen, 1966a, 1966b; Frederiksen, Saunders, & Wand, 1957). Known as the In-Basket Test, this technique has been adapted for testing Air Force officers in administrative positions, business executives, administrators in government agencies, and school principals. Simulating the familiar "in-basket" found on the administrator's desk, this test provides a carefully prepared set of incoming letters, memoranda, reports, papers to be signed, and similar items. Before taking the test, the examinee has an opportunity to study background materials for orientation and information regarding the hypothetical job. During the test proper, his task is to handle all the matters in his in-basket as he would on the job. All actions must be recorded in writing but may include letters, memos, decisions, plans, directives, information to be obtained or transmitted by telephone, agenda for meetings, or any other notes.

Simulators merge imperceptibly with *job samples*. In the latter, the task set for the examinee is actually a part of the work he is to perform on the job. Many such tests are custom-made to fit specific jobs. The representativeness of the work sample and the closeness with which the task duplicates actual job conditions are essential considerations. Familiar examples are tests for drivers, as well as several standardized tests for office skills, such as typing, stenography, bookkeeping, and the operation of various business machines. For certain types of jobs, written or oral information tests fall into the same category. It should be noted, of course, that such information tests are not substitutes for performance tests in jobs requiring manipulative skills. Moreover, they must be carefully checked to ensure that they do not require a higher level of verbal comprehension than is demanded by the job. In evaluating job-sample tests of any sort, the procedures of content validation are appropriate (see Ch. 6).

In this brief overview, we have considered the need for local validation of industrial tests, the difficulties in the way of obtaining meaningful criterion-related validity data, and several alternative solutions. Major features of these alternative solutions include a full and systematic job analysis and the identification and assessment of subcriteria, or separate facets of job performance. Progress toward these goals has been substantially advanced by recent research on the taxonomy of human performance (Fleishman, 1972; Theologus & Fleishman, 1971). This approach is also exemplified in the development of the Position Analysis Questionnaire (PAQ), a job-analysis instrument that permits the rating of jobs in terms of common behavioral requirements (McCormick, Jeanneret, & Mecham, 1972).

SHORT SCREENING TESTS FOR INDUSTRIAL PERSONNEL

In this and the following section, we shall examine typical published tests available for either industrial uses or occupational counseling. Several tests of general mental ability have been specially developed for the rapid, preliminary screening of industrial personnel. Some of these tests represent abridged versions of earlier tests; others have been specifically constructed for the purpose. In all these tests, administration and scoring are simplified, and an effort is usually made to give the content face validity in an industrial setting.

It should be clearly recognized that general screening tests may have fairly high validity for some jobs and little or no validity for others. The type of behavior sampled by these tests is undoubtedly far more relevant to some types of jobs than to others. Jobs cannot simply be put into a hierarchy in terms of the amount of "intelligence" required, because the *type* of intelligence needed for different jobs varies. For many occupations, especially those requiring mechanical skills, tests of special aptitudes will serve as better predictors of achievement than will the general intelligence tests. This point is not always adequately stressed in the test manuals. In fact, some test manuals tend to create the erroneous impression that the general screening test can be used to predict success in almost every type of industrial work.

An early test that has enjoyed wide popularity for personnel screening is the Otis Self-Administering Test of Mental Ability, which introduced the spiral-omnibus arrangement of test items for ease of administration (See Ch. 11). This test has been used in screening applicants for such varied jobs as those of clerks, calculating-machine operators, assembly-line workers, and foremen and other supervisory personnel. Dorcus and Jones (1950) cited 36 validation studies in which the Otis test was checked against an industrial criterion. Not all of these studies yielded significant validity coefficients, of course, but many of them did. In semiskilled jobs, for example, the Otis test correlated moderately well with success in learning the job and ease of initial adaptation, but not with subsequent job achievement. This would be expected for jobs that are largely routine, once they are learned. Also, for high-level professional personnel, who represent a select group in terms of academic achievement, correlations between Otis scores and criteria of job success were usually negligible, since this test does not discriminate adequately at the upper levels.

Adaptations of the original Otis test have also been prepared for industrial use, the best known being the Wonderlic Personnel Test. Having been widely used for many years, this test has accumulated extensive normative data on industrial samples. Despite its brevity, reliability appears to be adequate. Validity is less well substantiated. And the test has undoubtedly been applied beyond the type of jobs for which it is appropriate. Performance on the Wonderlic is closely related to amount of education completed. In jobs with only minimal educational requirements, therefore, this test may unfairly discriminate against educationally disadvantaged minority group members. In an effort to meet this problem in 1972, Wonderlic published a separate collection of black norms based on about 38,500 job applicants. These norms are reported for 81 common job titles and are subdivided with regard to amount of education (Grades 7 to 16), age (15 to 44 years), sex, and 10 geographical regions.

A somewhat different approach is illustrated by the Wesman Personnel Classification Test. Like most current intelligence tests, it yields Verbal, Numerical, and Total scores. The Verbal score is based on an 18-minute verbal analogies test, in which each item contains two blanks, as illustrated in Figure 79. The Numerical score is derived from a 10-minute arithmetic computation test whose items were designed so as to put a premium on ingenuity and ability to perceive numerical relations. Two parallel forms are available. Percentile norms on each of the three scores are reported for groups of students, job applicants, and employees, each group including from 93 to 1,476 cases. Parallel-form reliability coefficients for V, N, and Total scores fall mostly in the .80's. Correlations of V and N scores vary from .25 to .57, indicating that the overlap of the two parts is small enough to justify retention of separate scores.

Mean scores on the test show progressive rise with increasing educational and occupational level in the groups compared. Correlations with criteria of vocational success, usually based on supervisors' ratings, range from .29 to .62. From the nature of the items, as well as from the distribution of scores reported for various groups, it appears that this test may be better suited for higher-level than for lower-level personnel. It also seems likely that the predominantly academic content of the items would not hold the interest of lower-level job applicants and would lack face validity for them.

Each question in Part I is a sentence with the first word and the last word left out. You are to pick out words to fill in the blanks so that the sentence will be true and sensible.

For the first blank, pick out a **numbered** word—1, 2, 3, or 4. For the blank at the end of the sentence, pick out one of the **lettered** words —A, B, C, or D. Then write the number and the letter you have picked on the line at the right.

Example 1. ..... is to water as eat is to 2C

1. continue	2. drink	3. foot	4. girl
A. drive	B. enemy	C. food	D. industry

Drink is to water as eat is to **food.** **Drink** is numbered 2, and **food** is lettered C, so 2C has been written on the line at the right.

Fig. 79. Sample Analogies Item from the Wesman Personnel Classification Test.

(Reproduced by permission. Copyright 1946, The Psychological Corporation, New York, N.Y. All rights reserved.)

For the rapid screening of industrial personnel at the other end of the distribution, a more appropriate instrument is the Personnel Tests for Industry (PTI). This battery includes a 5-minute Verbal Test, a 20-minute Numerical Test, and a 15-minute Oral Directions Test. The three tests may be used together or separately. The Oral Directions Test, administered by phonograph or tape recording, is suitable for applicants with limited schooling or with a foreign language background. It is also available in a Spanish edition.

Alternate-form reliabilities of the Verbal and Numerical Tests range from .73 to .92; split-half reliabilities of the Oral Directions Test range from .82 to .94. Norms for all three tests are reported for various industrial and educational samples. Older persons may be somewhat handicapped on the Oral Directions Test because of its dependence on auditory discrimination and speed. The test is rather heavily weighted with perceptual and spatial items, and also depends to a considerable extent on immediate memory for auditory instructions. Available data suggest that it discriminates somewhat better at the lower intellectual levels and may be particularly useful in screening applicants for such jobs as general laborer, maintenance and service worker, and messenger. Reference may also be made to the adult basic literacy tests (Ch. 14), designed for similar populations but more specifically directed to the assessment of educational skills.

SPECIAL APTITUDE TESTS

Even prior to the development of multiple aptitude batteries, it was generally recognized that intelligence tests were limited in their coverage of abilities. Efforts were soon made to fill the major gaps by means of special aptitude tests. Among the earliest were those designed to measure mechanical aptitude. Since intelligence tests concentrate chiefly on "abstract" functions involving the use of verbal or numerical symbols, a particular need was felt for tests covering the more "concrete" or "practical" abilities. Mechanical aptitude tests were developed partly to meet this need.

The demands of vocational selection and counseling likewise stimulated the development of tests to measure mechanical, clerical, musical, and artistic aptitudes. Tests of vision, hearing, and motor dexterity have also found their principal applications in the selection and classification of personnel for industrial and military purposes. It is thus apparent that a strong impetus to the construction of all special aptitude tests has been provided by the urgent problems of matching job requirements with the specific pattern of abilities characterizing each individual.

A word should be added about the concept of special aptitudes. The term originated at a time when the major emphasis in testing was placed on general intelligence. Mechanical, musical, and other special aptitudes were thus regarded as supplementary to the "IQ" in the description of the individual. With the advent of factor analysis, however, it was gradually recognized that intelligence itself comprises a number of relatively independent aptitudes, such as verbal comprehension, numerical reasoning, numerical computation, spatial visualization, associative memory, and the like. Moreover, several of the traditional special aptitudes, such as mechanical and clerical, are now incorporated in some of the multiple aptitude batteries.

What, then, is the role of special aptitude tests? First, there are certain areas, such as vision, hearing, motor dexterity, and artistic talents, that are rarely included in multiple aptitude batteries. The situations requiring tests in these areas are too specialized to justify the inclusion of such tests in standard batteries. Special aptitude tests are also employed, however, in areas that are covered in multiple aptitude batteries, such as clerical and mechanical aptitudes. In several testing programs, tests of general intelligence are combined with specially selected tests of other relevant aptitudes. One reason for this practice is to be found in the extensive normative and validation data available for some widely used tests of special aptitudes. Another reason is undoubtedly the flexibility that this procedure provides, not only in the choice of relevant aptitudes but also in the fullness with which each aptitude is measured for specific purposes.

MOTOR FUNCTIONS. Many tests have been devised to measure speed, coordination, and other characteristics of movement responses. Most are concerned with manual dexterity, but a few involve leg or foot movements that may be required in performing specific jobs. Some measure a combination of motor and perceptual, spatial, or mechanical aptitudes, thus overlapping the tests to be discussed in the next sections. Motor tests are characteristically apparatus tests, although several paper-and-pencil adaptations have been designed for group administration. Examples can be found in the General Aptitude Test Battery, discussed in Chapter 13. Some of these printed motor tests may prove valid in their own right as predictors of practical criteria. Available evidence indicates, however, that there is little or no correlation between printed tests and apparatus tests designed to measure the same motor functions (Fleishman, 1958; Melton, 1947).

The principal application of motor tests has been in the selection of industrial and military personnel. Frequently, they have been custom-made to meet the requirements of specific jobs. Many are constructed on the principle of simulation, or "job miniature." This means that the test closely reproduces all or part of the movements required in the performance of the job itself. These tests are not miniatures in the sense of involving smaller movements. On the contrary, to ensure validity, the test and the job should call for the use of the same muscle groups. Many examples of motor coordination tests, ranging from simple to very complex, are provided by the classification testing program conducted by the Air Force during World War II (Melton, 1947). While these tests are not commercially available, their published report represents a rich source of test material for the research worker and the test constructor.

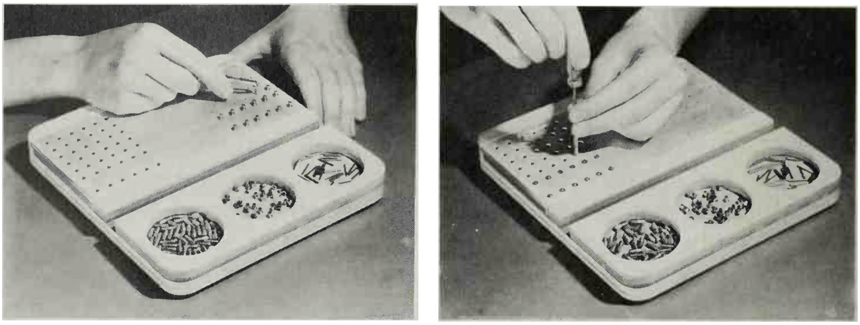


FIG. 80. Crawford Small Parts Dexterity Test.
(Courtesy The Psychological Corporation.)

A test requiring several simple manipulative skills is the Crawford Small Parts Dexterity Test, shown in Figure 80. In Part I of this test, the examinee uses tweezers to insert pins in close-fitting holes, and then places a small collar over each pin. In Part II, small screws are placed in threaded holes and screwed down with a screwdriver. The score is the time required to complete each part. Split-half reliability coefficients between .80 and .95 are reported for the two parts of this test. Despite the apparent similarity of the functions required by Parts I and II, the correlations between the two parts ranged from .10 to .50 in several industrial and high school samples, with a median correlation of .42.

Another widely used manual dexterity test which, however, utilizes no tools is the Purdue Pegboard. This test is said to provide a measure of two types of activity, one requiring gross movements of hands, fingers, and arms, and the other involving tip-of-the-finger dexterity needed in small assembly work. First, pins are inserted individually in small holes with the right hand, left hand, and both hands together, in successive trials. In another part of the test, pins, collars, and washers are assembled in each hole. The prescribed procedure for this activity involves the simultaneous use of both hands.

What can be said about the effectiveness of motor tests as a whole? The most important point to note in evaluating such tests is the high degree of *specificity* of motor functions. Intercorrelations and factor analyses of large numbers of motor tests have failed to reveal broad group factors such as those found for intellectual functions. In extensive factor-analytic research, Fleishman and his associates (Fleishman, 1954, 1958, 1972; Fleishman & Ellison, 1962; Fleishman & Hempel, 1956) identified the following major factors:

Multi-Limb Coordination: Ability to coordinate gross movements requiring the simultaneous use of more than one limb in operating controls.

Control Precision: Ability to make highly controlled and precise muscular adjustments—important in the rapid and accurate operation of controls by hand, arm, and foot movements.

Response Orientation: Ability to select the appropriate response under highly speeded conditions—identified in complex coordination tests in which each pattern of signals requires a different choice of controls and direction of movement.

Reaction Time: Speed with which an individual is able to respond to a stimulus when it appears—found to be independent of specific response required and of whether the stimulus is auditory or visual.

Speed of Arm Movement: Speed with which gross arm movements can be made, regardless of precision.

Rate Control: Ability to make continuous anticipatory motor adjustments relative to changes in speed and direction of a moving target—the common factor in pursuit and tracking tests.

Manual Dexterity: Ability to make skillful, well-controlled arm-hand movements in manipulating fairly large objects under speed conditions.

Finger Dexterity: Ability to make skillfull, controlled manipulations of small objects, involving primarily finger movements.

Arm-Hand Steadiness: Ability to make precise arm-hand positioning movements where strength and speed are minimized.

Wrist-Finger Speed: Traditionally called "tapping," this ability is best measured by paper-and-pencil tests requiring rapid tapping of the pencil in relatively large areas.

Aiming: A narrowly defined ability measured chiefly by paper-and-pencil dotting tests which require subject to place a dot accurately and rapidly in each of a series of small circles.

Still other factors have been identified in the area of gross bodily movements, as exemplified by tests of physical fitness (Fleishman, 1964). These include: static strength (continued exertion of force up to a maximum), dynamic strength (muscular endurance requiring repeated exertion of force), explosive strength (mobilization of energy for bursts of muscular effort, as in sprints and jumps), trunk strength, extent flexibility, dynamic flexibility (repeated, rapid flexing movements), body coordination, gross body equilibrium, and stamina (capacity to sustain maximum effort).

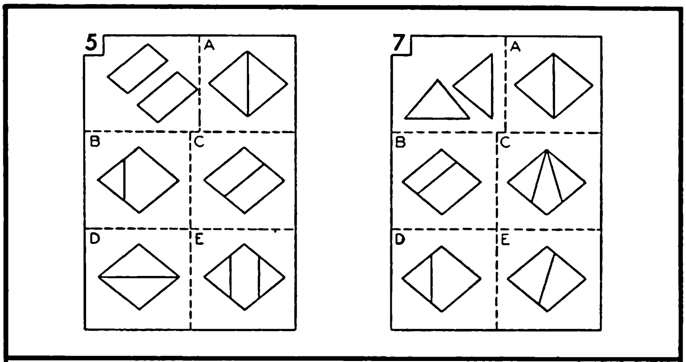
There is evidence that the abilities called into play by motor tests may change with *practice.* Fleishman (1962) factor analyzed the scores obtained during continued practice on the Air Force Complex Coordination Test, a test reproducing many of the eye-hand-foot coordinations required in piloting a plane. Both number and nature of factors identified at different stages of practice varied. At the early stages, nonmotor factors such as spatial orientation, visualization, mechanical experience, and perceptual speed entered into performance together with motor factors. With increasing practice, the importance of the intellectual factors declined and that of the motor factors increased. In the last stages of practice, the only common factors having significant weights were speed of arm movement and control precision. A factor specific to the Complex Coordination Test also increased in weight with practice. Similar results were obtained in later analyses of other complex motor tests.

In considering the *validity* of motor tests, we need to differentiate between complex motor tests that closely resemble the particular criterion performance they are trying to predict and tests of simple motor functions designed for more general use. The former are well illustrated by some of the Air Force tests. Such complex, custom-made tests that reproduce the combination of motor aptitudes required by the criterion have shown fair validity. The Complex Coordination Test of the Air Force, for example, considerably improved the prediction of performance in pilot training. For most purposes, however, the use of such tests is not practicable, since a very large number of tests would have to be devised to match different criteria. With regard to commercially available motor tests, the functions they measure are very simple and their validities against most criteria are not high. For this reason, such tests can serve best as part of a selection battery, rather than as single predictors. In general, motor tests have been most successful in predicting performance on routine assembling and machine-operating jobs (Ghiselli, 1966). As the jobs become less repetitive, perceptual and intellectual factors come to play a more important part.

MECHANICAL APTITUDES. Mechanical aptitude tests cover a variety of functions. Motor factors enter into some of the tests in this category, either because the rapid manipulation of materials is required in the performance of the test, or because special subtests designed to measure motor dexterity are included in a paper-and-pencil test. In terms of the factors discussed in Chapter 13, perceptual and spatial aptitudes play an important part in many of these tests. Finally, mechanical reasoning and sheer mechanical information predominate in a number of mechanical aptitude tests.

It is important to recognize the diversity of functions subsumed under the heading of mechanical aptitude, since each function may be differently related to other variables. For example, mechanical information tests are much more dependent on past experience with mechanical objects than are abstract spatial or perceptual tests. Similarly, sex differences may be reversed from one of these functions to another. Thus, in manual dexterity and in perceptual discrimination tests, women generally excel; in abstract spatial tests, a small but significant average difference in favor of males is usually found; while in mechanical reasoning or information tests, men are markedly superior, the difference increasing with age (Anastasi, 1958a, Ch. 14).

Among the aptitudes included in all multiple aptitude batteries is spatial aptitude. This is the ability measured by the Space Relations test of the DAT; it corresponds to the S factor in Thurstone's primary mental abilities and to the *k* factor in the British hierarchical model of intelligence. It has also been found to have a high loading in many performance and nonlanguage tests of general intelligence. One of the best single measures of this aptitude is the Revised Minnesota Paper Form Board Test. Available in two equivalent forms, this test was originally developed as a paper-and-pencil adaptation of the familiar formboard type of test.



For each item, the subject must choose the figure that would result if the pieces in the first section were assembled.

FIG. 81. Sample Items from the Revised Minnesota Paper Form Board.

(Reproduced by permission. Copyright 1941, renewed 1969 by The Psychological Corporation, New York, N.Y. All rights reserved.)

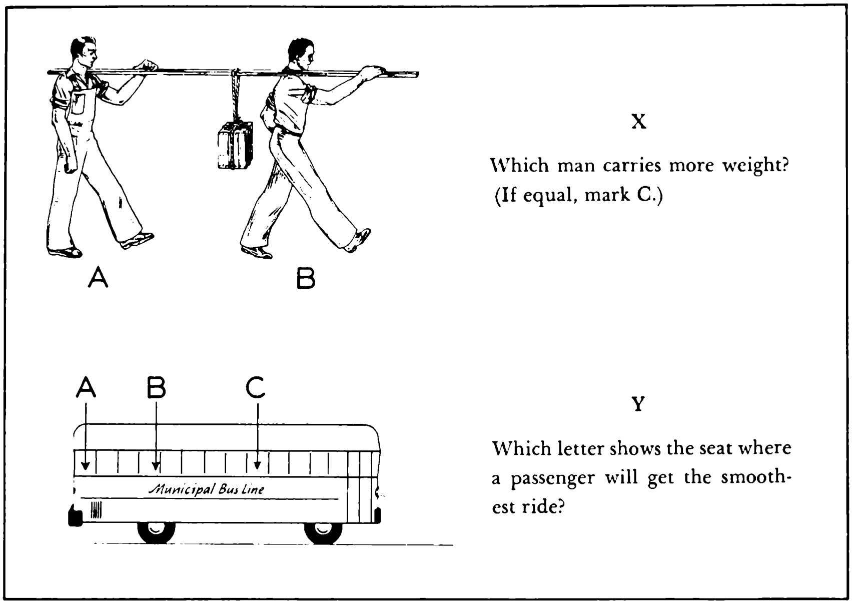
Two sample items are reproduced in Figure 81. Each item consists of a figure cut into two or more parts. The examinee must determine how the pieces would fit together into the complete figure; he then chooses the drawing that correctly shows this arrangement.

The manual for the Minnesota Paper Form Board provides extensive tables of norms based on large and well-defined educational and industrial samples. Alternate-form reliability coefficients are in the .80's. An unusually large number of studies have been conducted with this test. The results indicate that it is one of the most valid available instruments for measuring the ability to visualize and manipulate objects in space. Among the criteria employed in this research were performance in shop courses, grades in engineering and in other technical and mechanical courses, supervisors' ratings, and objective production records. The test has also shown some validity in predicting the achievement of dentistry and art students.

Another major type of mechanical aptitude test is concerned with mechanical information, mechanical reasoning, or mechanical comprehension. While requiring some familiarity with common tools and mechanical relations, these tests assume no more technical knowledge than can be acquired through everyday experience in an industrial society such as ours. Some of the early tests in this field required the examinee to assemble common mechanical objects from the given parts. For general testing purposes, paper-and-pencil group tests are now widely employed.

A well-known example of this type of test is the Bennett Mechanical Comprehension Test. Utilizing pictures about which short questions are to be answered, this test emphasizes the understanding of mechanical principles as applied to a wide variety of everyday-life situations. Two items employed in the fore-exercise of this test are reproduced in Figure 82. This test has been used widely for both military and civilian purposes. In its current revision, it is available in two equivalent forms, S and T. Covering a wide difficulty range, these forms are suitable for such groups as high school students, industrial and mechanical job applicants and employees, and candidates for engineering school. Percentile norms are provided for several groups, permitting the user to choose the most appropriate group in terms of educational level, specialized training, or prospective job category. Odd-even reliability coefficients of each form range from .81 to .93 in different groups.

With regard to validity, the manual gives some thirty correlations with training or job-performance criteria, most of which were obtained with earlier forms of the test. While varying widely, about two thirds are above .30 and extend to the .60's. Other published studies with both civilian and military forms provide good evidence of both concurrent and predictive validity for mechanical trades and engineering (Ghiselli,



X

Which man carries more weight?
(If equal, mark C.)

Y

Which letter shows the seat where a passenger will get the smoothest ride?

FIG. 82. Sample Items from the Bennett Mechanical Comprehension Test.

(Reproduced by permission. Copyright 1940, renewed 1967; 1941, renewed 1969; 1942, renewed 1969; © 1967, 1968 by The Psychological Corporation, New York, N.Y. All rights reserved.)

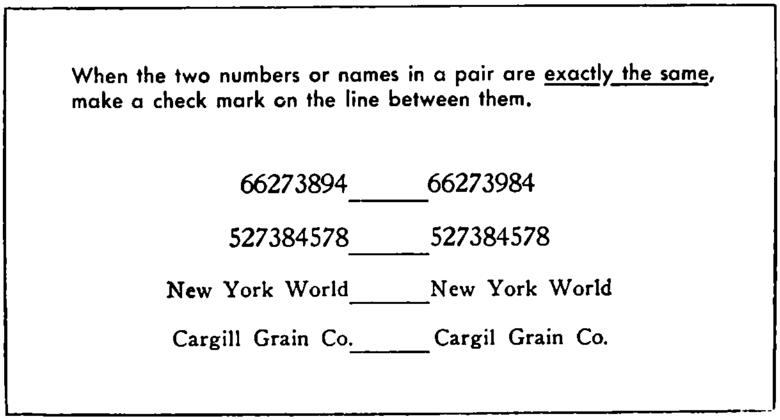
1966; Patterson, 1956). During World War II, this test proved to be one of the best predictors of pilot success (Guilford & Lacey, 1947, p. 843). Its validity for this purpose seems to have resulted chiefly from the contribution of a mechanical information factor and a spatial visualization factor, which together accounted for about 60 percent of the variance of its scores (Guilford & Lacey, 1947, pp. 336–339).

For examinees with limited reading skills, Forms S and T may be administered with tape recorded instructions and questions. Spanish versions are also available for both regular and oral administration. In addition, there is a more difficult Form CC, permitting finer discrimination at higher ability levels. Finally, it should be added that the Mechanical Reasoning Test of the Differential Aptitude Tests (Ch. 13) is sufficiently similar to the Bennett Mechanical Comprehension Test to be regarded as another form of the same test.

CLERICAL APTITUDES. Tests designed to measure clerical aptitudes are characterized by a common emphasis on perceptual speed. A well-known example is the Minnesota Clerical Test, which consists of two separately timed subtests, Number Comparison and Name Comparison. In the first, the examinee is given 200 pairs of numbers, each containing from 3 to 12 digits. If the two numbers in the pair are identical, he places a check mark between them. The task is similar in the second subtest, proper names being substituted for numbers. Sample items from both subtests are reproduced in Figure 83. Although a deduction is made for errors, the scores depend predominantly on speed. Performance on this test is influenced by the examinee's response set. Occasionally a very careful worker will obtain a poor score because he proceeds slowly in order to avoid errors. By contrast, the examinee who emphasizes speed at the expense of accuracy will complete many more items and will be penalized only a few points as a result of errors. The possible effect of response set should therefore be taken into account when interpreting a low score, especially when such a score is obtained by an otherwise able or promising person.

Percentile norms on the Minnesota Clerical Test are reported for several large samples of clerical applicants and employed clerical workers, as well as for boys and girls in Grades 8 to 12. Retest reliability coefficients over intervals of several months fall mostly in the .70's and .80's. These values are probably underestimates, since the intervals were longer than those usually employed for estimating reliability.

Data on both concurrent and predictive validity of the Minnesota Clerical Test are provided by a number of independent studies. Moderately high correlations have been found between scores on this test and ratings by office supervisors or by commercial teachers, as well as performance records in courses and in various kinds of clerical jobs. Several studies employed the method of contrasted groups. Thus, comparisons



When the two numbers or names in a pair are exactly the same, make a check mark on the line between them.

66273894______66273984

527384578______527384578

New York World______New York World

Cargill Grain Co.______Cargil Grain Co.

FIG. 83. Sample Items from the Minnesota Clerical Test.

(Reproduced by permission. Copyright 1933, renewed 1961 by The Psychological Corporation, New York, N.Y. All rights reserved.)

are reported between different levels of clerks, between clerical workers and persons engaged in other occupations, and between employed and unemployed clerks. All these comparisons yielded significant differences in mean scores in the expected direction. A marked and consistent sex difference in favor of women has been found on this test, beginning in childhood and continuing into adulthood.

It is apparent, of course, that such a relatively homogeneous test as the Minnesota Clerical Test measures only one aspect of clerical work. Clerical jobs cover a multiplicity of functions. Moreover, the number and particular combination of duties vary tremendously with the type and level of job. Despite such a diversity of activities, however, job analyses of general clerical work indicate that a relatively large proportion of time is spent in such tasks as classifying, sorting, checking, collating and stapling, stuffing and sealing envelopes, and the like (Bennett & Cruikshank, 1949). Speed and accuracy in perceiving details would thus seem to be of primary importance for the clerical worker, especially in the lower-level, routine jobs (Hay, 1954).

To be sure, many types of jobs besides that of clerk require perceptual speed and accuracy. Inspectors, checkers, packers, and a host of other factory workers obviously need this ability. It is interesting to note in this connection that the Minnesota Clerical Test has also been found to have some validity in predicting the performance of such workers (Dorcus & Jones, 1950). However, it is likely that higher validity may be obtained in such cases by designing a similar test with pictorial rather than with verbal or numerical material. It will be recalled that, in the GATB prepared by the United States Employment Service, separate tests are included for the *Q* and *P* factors. The *Q* factor appeared in number- and word-checking tests similar to those that make up the Minnesota Clerical Test. The *P* factor, on the other hand, occurred in tests requiring the perception of similarities and differences in spatial items, and is probably more closely related to the inspection of materials.

Several tests of clerical aptitude combine perceptual speed and accuracy with other functions required for clerical work. Among the measures used for the latter functions are "job sample" types of tests for such activities as alphabetizing, classifying, coding, and the like. In addition, some measure of verbal and numerical ability may be included, to serve in lieu of a general intelligence test. Other clerical aptitude tests include such office skills as business vocabulary, business information, spelling, and language usage.

ARTISTIC APTITUDES. Successful achievement in the visual arts calls for a multiplicity of aptitudes and personality traits. Different combinations of these characteristics may be required by diverse specialties and art forms. The sculptor and the painter utilize different skills; among painters, the traditional portraitist and the surrealist artist undoubtedly differ in a number of significant ways. In pursuing a career in art, moreover, an individual may play any of several roles, each with its distinct set of qualifications. Thus, the creative artist, art teacher, critic, dealer, and museum administrator approach art with different patterns of skills, knowledge, interests, attitudes, and feelings.

Standardized tests of artistic aptitudes have traditionally concentrated on a few aptitudes considered to be basic to artistic activity, whatever its form. Among these tests, however, a distinction can be made between those concerned only with aesthetic appreciation and those measuring also productive skills. Obviously a person may be a highly discriminating and sophisticated connoisseur of paintings without himself being able to paint. But artistic production, except at a routine and mechanical level, undoubtedly presupposes superiority in both appreciative and productive skills. Thus, tests of art appreciation have a broader applicability than tests of production.

Tests of *artistic appreciation* have generally followed a common pattern. In each item, the examinee is requested to express his preference regarding two or more variants of the same object. One variant is either an original by an eminent artist or a version preferred by the majority of a group of art experts. The other versions represent deliberate distortions designed to violate some accepted principle of art. Any controversial items, on which a clear consensus of experts cannot be obtained, are normally eliminated from such a test.

In the development of art appreciation tests, both original item selection and subsequent validation procedures depend heavily on the opinions of contemporary art experts within our culture. It is well to bear this fact in mind when interpreting scores. Essentially, such tests indicate the degree to which the individual's aesthetic taste agrees with that of contemporary art experts. The representativeness of the particular group of experts employed in developing the test is obviously an important consideration. Insofar as aesthetic standards or taste may change with time, the periodic rechecking of scoring key and item validities is likewise desirable.

A well-established and unusually viable test in the field of art appreciation is the Meier Art Judgment Test. First published in 1929 and revised in 1940, this test employs relatively timeless art works, that will not readily go out of date. Most are paintings or drawings by acknowledged masters, while a few represent vases or designs suitable for pottery. All reproductions are in black and white. Each item contains only two versions, an original and a variation in which the symmetry, balance, unity, or rhythm has been altered. This test thus concentrates on the judgment of aesthetic organization, generally considered to be a key factor in artistic sensitivity. In order to rule out the contribution of perceptual accuracy, the examinee is told in what detail the two versions of each picture differ. An illustrative item is reproduced in Figure 84.

Percentile norms are given for junior and senior high school students and for adults. All norms are derived largely from students in art courses, whether in high school, college, or special art schools. Data were gathered in 25 schools scattered throughout the United States. Split-half reliability coefficients between .70 and .84 are reported for relatively homogeneous samples.

Most of the available evidence regarding validity of the Meier Art Judgment Test can be summarized under the headings of item selection and contrasted group performance, although a few correlations with independent criteria of artistic accomplishment have also been reported. Items were selected on the basis of several criteria, including: original art work versus deliberate distortion; agreement among a group of art

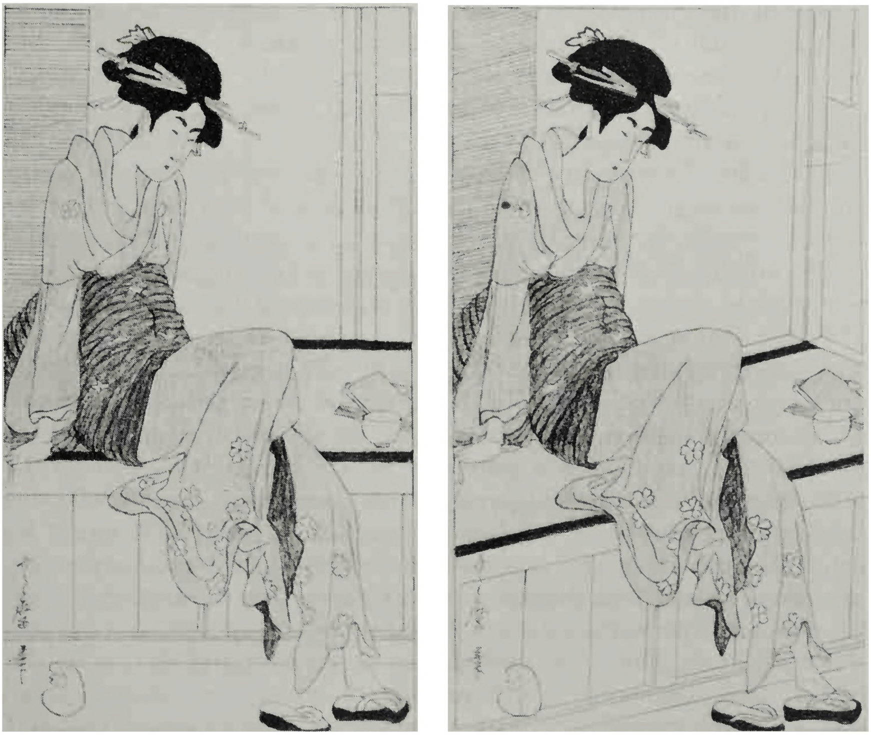


Fig. 84. Item Similar to Those Employed in the Meier Art Judgment Test. The difference between the two versions is in the angle at which the window seat crosses the picture.

(Courtesy Norman C. Meier.)

experts; choice of original version by large majority of a miscellaneous sample of about 1,000 persons; and internal consistency, i.e., correspondence between item choice and total score.

Performance on the Meier Art Judgment Test exhibits a sharp differentiation in terms of age, grade, and art training. Thus, art faculty score higher than nonart faculty, art students higher than comparable nonart students. The extent to which these group differences result from selection or from previous art training cannot be determined from available data. Although no validity coefficients are given in the manual, a few are reported in other published sources. Correlations ranging from .40 to .69 have been found between scores on this test and art grades or ratings of artistic ability (H. A. Carroll, 1933; Kinter, 1933; Morrow, 1938).

As in most artistic aptitude tests, the Meier Art Judgment Test has regularly shown negligible correlation with traditional intelligence tests, such as the Stanford-Binet or group verbal tests. This does not mean, however, that abstract intelligence or scholastic aptitude is unrelated to success in an art career. In fact, research by Meier and his associates suggested that, for higher levels of artistic accomplishment, superior scholastic aptitude is a decided asset. In one study, for example, the mean IQ of successful artists was found to be 119 (Tiebout & Meier, 1936). Similarly, a group of artistically gifted children had IQ's ranging from 111 to 166 (Meier, 1942).

Most tests of *productive ability* in the visual arts are actually work-samples. As such, they are undoubtedly influenced to a large extent by formal art training. A few tests, however, have been designed specially for the prediction of performance in subsequent training and have therefore tried to minimize the demands on technical skills and information. The Horn Art Aptitude Inventory may be cited as an example of this type of test. Concerned exclusively with the production of original drawings, this test has a fairly high ceiling and shows adequate discrimination among applicants for admission to art schools. In a major part of the test, the examinee is required to sketch a picture in each of 12 rectangles, building on a few given lines. A sample item, used for demonstration purposes in the test, is reproduced in Figure 85. It might be noted that this procedure is sometimes employed in tests of creativity, as in the Torrance tests described in Chapter 13. In such tests, however, the examinee's performance is not scored for artistic quality.

Scoring of the Horn Art Aptitude Inventory illustrates the use of the product-scale technique, which provides a criterion-referenced rather than a norm-referenced evaluation. Samples of excellent, average, and poor work are included in the manual as a basis for rating the examinee's drawings. As an additional scoring guide, the manual lists certain factors to be considered, such as order, clarity of thought and presentation, quality of line, use of shading, fertility of imagination, and scope of in-

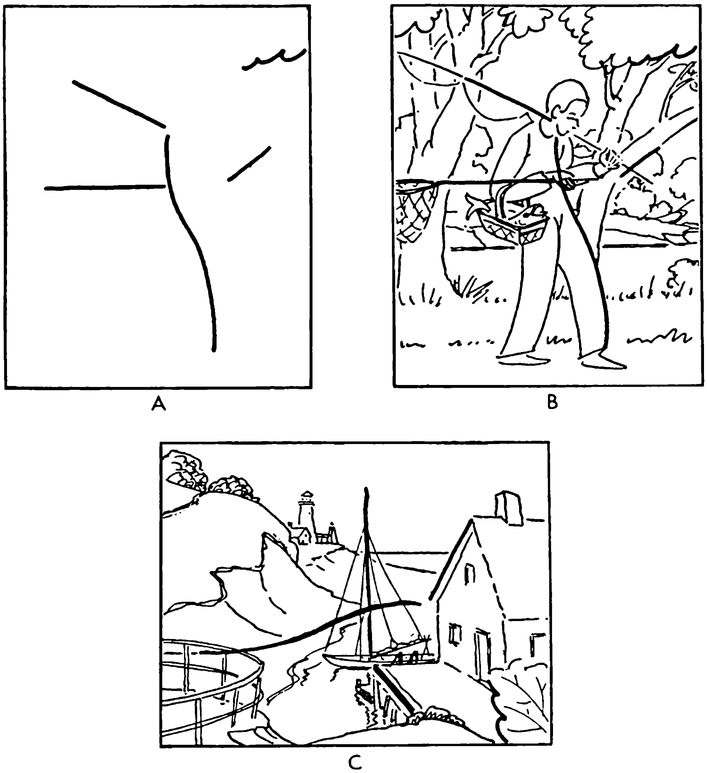


Fig. 85. Sample Item from the "Imagery" Test of the Horn Art Aptitude Inventory. The first rectangle shows the stimulus lines; the other two contain drawings made with these stimulus lines. In the second drawing, the card has been turned to a horizontal position.

(From Horn & Smith, 1945, p. 351; reproduced by permission of American Psychological Association.)

terests. Although the scoring still leaves much to subjective judgment, correlations of .79 to .86 are reported between the results obtained by different scorers. No systematic validation research was undertaken with this test; but two studies reported in the manual show promising evidence of both concurrent and predictive validity.

MUSICAL APTITUDES. During the first four decades of this century, extensive research on the psychology of music was conducted at the University of Iowa under the direction of Carl E. Seashore (1938). One of the outcomes of these investigations was the preparation of the Seashore Measures of Musical Talents. In its present form, this series consists of six tests covering pitch, loudness, rhythm, time, timbre, and tonal memory. The Seashore tests are now available on both an LP record and a tape recording. Either or both of these modes of presentation are utilized in most current musical aptitude tests to permit group administration and to ensure uniformity of procedure.

Each item in the Seashore tests consists of a pair of tones or tonal sequences. In the pitch test, the examinee indicates whether the second tone is higher or lower than the first. The items are made progressively more difficult by decreasing the difference in pitch between the two tones in each pair. In the loudness test, the question is whether the second tone is stronger or weaker than the first. The rhythm test requires the comparison of rhythmic patterns that are either the same or different within each pair. In the time test, the examinee records whether the second tone in each pair is longer or shorter than the first. The timbre test calls for the discrimination of tone quality, the two tones in each pair being either the same or different in this respect. In the tonal memory test, short series of three to five tones are played twice in immediate succession. During the second playing, one note is changed, and the examinee must write the number of the altered note (i.e., first, second, etc.).

The Seashore tests are applicable from the fourth grade to the adult level. The testing of younger children by this procedure has not proved feasible because of the difficulty of sustaining interest and attention. Even above the age of 10, the scores on these tests may be lowered by inattention. The scores are not combined into a single total, but are evaluated separately in terms of percentile norms. These norms are reported for Grades 4 to 5, 6 to 8, and 9 to 16, the normative samples for each test and grade level containing from 377 to 4,319 cases. Age changes are slight and sex differences are negligible. The tests are probably somewhat susceptible to practice and training, although studies of these effects have yielded conflicting results.

Kuder-Richardson reliability coefficients of the six Seashore tests range from .55 to .85 within the three normative grade-level groups. Only content validity is discussed in the manual, Seashore having argued over the years that this is the most appropriate type of validity for such tests. It is undoubtedly true that ability to discriminate pitch, loudness, timbre, and other properties of tones is essential to both appreciation and production of music. To predict musical achievement, however, we need to know much more. What is the minimum cutoff point for different kinds of musical activities? Is there any correlation between test scores and musical performance beyond the cutoff point? What is the relative importance of each of the functions measured by the tests, both in relation to each other and in relation to the entire array of requisite traits? A few scattered studies provide some evidence of predictive validity against various criteria of performance in music training. Many of these validity coefficients are low, few reaching .30 to .40. Apart from the unreliability of criterion measures and the complexity of factors affecting musical achievement, it should be noted that these investigations were conducted on selected samples of music students. Presumably most of the individuals falling below a certain minimum in relevant auditory capacities had already been eliminated from such groups.

Correlations with intelligence tests are negligible, as would be expected for special aptitude tests. Intercorrelations among the six tests are higher than had been anticipated. The functions measured by the different tests are thus less independent than had originally been supposed, a fact that has also been confirmed by factor analysis (McLeish, 1950). It should also be noted that the Seashore tests, or adaptations of them, have proved helpful in predicting achievement in certain civilian and military specialties requiring auditory discrimination, such as those of sonar operator and radiotelegrapher.

A more comprehensive approach to the measurement of musical aptitude was followed in a battery developed in England by Wing (1941, 1962). Applicable from the age of 8 on, the Wing tests depart from the "atomistic" sensory orientation of the Seashore tests and make use of musically meaningful content. Piano music is employed in each of the seven tests, which cover chord analysis, pitch discrimination, memory for pitch, harmony, intensity, rhythm, and phrasing. The first three tests require sensory discrimination, but at a somewhat more complex level than in the Seashore tests. In the other four, the examinee compares the aesthetic quality of two versions. Thus the battery places considerable emphasis on music appreciation.

Presented on tape, the entire battery requires approximately one hour but may be administered in two sessions. Norms are provided for total scores on the seven-test battery, as well as for a subtotal on the first three tests only. It is suggested that the first three tests may be used as a short battery with younger children or as a preliminary screening instrument. Based on a standardization sample of over 8,000 cases, the norms are reported separately for each age from 8 to 17 (adult). For older children and adults, both retest and split-half reliabilities of total scores on the entire battery are in the .90's; but they drop to the .70's for younger children.

Preliminary studies of validity in small groups yielded correlations of .60 or higher with teachers' ratings of musical ability. Other validation studies on both children and adults give promising evidence of predictive validity in terms of performance in music training. The use of total scores is supported by the identification of a general factor of musical ability in factorial analyses of music tests (McLeish, 1950; Wing, 1941). This factor, described as the cognitive aspect of musical ability, accounted for 30 to 40 percent of the total test variances. The Wing tests have high ceilings and may find their greatest usefulness in the selection of musically talented children for further training.

A similar approach was followed by E. Gordon (1965, 1967) in the more recently developed Musical Aptitude Profile. The taped selections in this test use exclusively violin and cello, rather than piano as in the Wing tests. Some reviewers have objected that this choice of instruments makes the test monotonous for the listeners. The entire battery comprises seven tests, grouped under tonal imagery (two tests), rhythm imagery (two tests), and musical sensitivity (three tests). Separate scores may be found in the seven tests, the three sections, and the total battery. The first four tests require perception of similarity or difference between pairs of phrases, with regard to melody, harmony, tempo, and meter, respectively. The last three call for musical preference within each pair of selections; the selections differ in phrasing, balance, or style. Insofar as they involve aesthetic judgment or taste, these three tests resemble the last four of the Wing tests.

In test-construction procedures, the Gordon tests compare favorably with other music tests. Percentile norms for Grades 4 to 12 are reported for the seven tests, the three part scores, and the composite. These norms were derived from over 12,800 pupils enrolled in a national sample of school systems, stratified with respect to geographical region, community size, and socioeconomic level. Additional norms are provided for musically select groups within the standardization sample. Reliability is satisfactory: in the .70's for separate tests, in the .80's for the three part scores, and in the .90's for composite battery scores. Although the tests have not been in use long enough to permit extensive accumulation of research results, available validity data are encouraging. Concurrent validity coefficients cited in the manual, with teachers' ratings as criteria, range from the .70's to the .90's for total scores. In a later, well-designed longitudinal study of predictive validity, the battery was administered to fourth- and fifth-grade children and the scores were compared with their performance at the end of one, two, and three years of instrumental training (Gordon, 1967). Correlations with objective evaluations of performance ranged from .53 to .71; correlations with teachers' ratings were in the .30's. Predictive validity did not decline—and in fact tended to rise—as the interval between test and criterion performance increased.[1]

TESTING IN THE PROFESSIONS

Standardized tests are used in many large-scale programs for the selection of professional personnel. Some of these programs are directed toward the selection of students for admission to professional training. Tests are administered to candidates for schools of medicine, dentistry, nursing, law, business, engineering, theology, architecture, and other professional fields. Although such testing programs emphasize aptitudes and the prediction of subsequent performance in specialized training, achievement tests on preprofessional courses constitute an important part of most batteries.

---

[1] For fuller coverage of available music tests, see Lehman (1968), as well as the appropriate section of the latest *Mental Measurements Yearbook.*

It should also be noted that in the selection of students for professional schools what is involved is not so much new types of tests as specially administered testing programs. There is no evidence that the various professional fields require any special aptitudes not already covered by available tests. The typical professional-school testing program includes a test of scholastic aptitude or general intelligence, one or more achievement tests on preprofessional training, and possibly tests of interests or other personality traits. Test results are often supplemented with biographical data, letters of recommendation, previous academic record, and interview ratings.

The intelligence test employed in such a program may be a standard scholastic aptitude test at an advanced level. More often it is specially designed so that the content can be slanted toward the particular professional field under consideration. Such a choice of content increases face validity, in addition to permitting better security control of test materials. There is also some evidence to suggest that the predictive validity of these special tests is a little higher than that of the intelligence or scholastic aptitude tests available for general use. The specialized scholastic aptitude tests often contain measures of reading comprehension for material similar to that which the student will encounter in professional school. Some of the tests yield separate verbal and quantitative scores. Spatial, mechanical, and motor aptitudes may also be separately tested when relevant to the field.

Another level at which standardized testing programs are making major inroads is that of specialty certification and the selection of job applicants following completion of training. Understandably these terminal testing programs draw much more heavily on achievement tests in specialized content areas; but more general types of tests are not excluded. Examples of major testing programs at this level include the National Teacher Examinations, constructed and administered by Educational Testing Service, and the NLN Achievement Tests in Nursing developed by the National League for Nursing. Educational Testing Service also works with several medical specialty boards and with the American Nurse's Association and other professional organizations in the development of certification examinations in the health professions. In the following sections, we shall look briefly at two programs illustrating the use of tests at pretraining and posttraining levels, respectively.

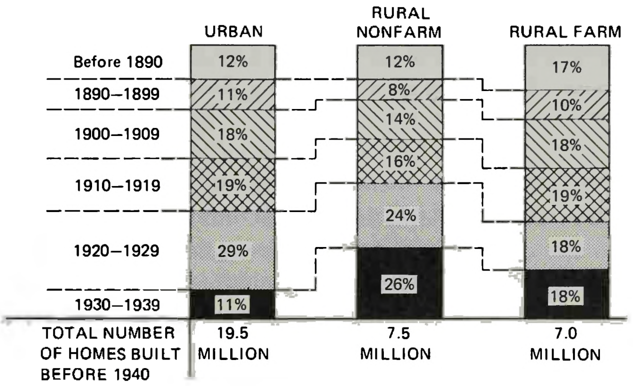
LAW SCHOOL ADMISSION TEST. Since 1948, the Law School Admission Test has been administered to law school candidates on a national basis. In its present form, this test includes seven separately timed sections and yields two scores, designated as the LSAT score and a Writing Ability score. The item types yielding the LSAT score vary somewhat from form to form but could include such items as reading comprehension, reading recall, principles and cases, and data interpretation. Examples of all item types can be found in the Law School Admission Bulletin, distributed to candidates.

The reading comprehension and reading recall items generally sample a wide range of content, using passages from the humanities, natural sciences, and social sciences. In each principles-and-cases item, a hypothetical case is briefly presented, together with a set of legal principles which may be real or imaginary. Following each principle are four alternatives relating to the possible applications of that principle to the stated facts. The candidate is asked to select the most reasonable alternative in each set. No prior legal knowledge is assumed. The data interpretation items, illustrated in Figure 86, are designed to measure comprehension of quantitative data presented in tabular or graphic form. The only nonverbal item types in the battery, these items nevertheless require careful reading of rather intricate verbal questions. In view of the prime importance of verbal comprehension in legal practice, it is understandable that legal aptitude tests have always been predominantly verbal in content.

The predictive validity of LSAT has been investigated in a continuing research program involving the joint participation of many law schools. Available data indicate that, when combined with prelaw grades, this test yields correlations with law school grades that cluster between .40 and .50. In one survey of 4,138 students in 25 law schools, first-year law school grades correlated .36 with undergraduate grades, .45 with LSAT scores, and .54 with the best weighted combination of LSAT scores and undergraduate grades (*LSAT Handbook*, 1964, p. 47). Despite the increasing homogeneity of student populations from the first to the third year of law school, LSAT scores tend to correlate as highly with three-year grades as with first-year grades (Breslow, 1957; Burnham & Crawford, 1957). It should nevertheless be borne in mind that validity coefficients vary widely from one law school to another, a finding that highlights the need for local validation. Moreover, the correlations between test scores and grades will decrease as students are more stringently selected on the basis of test scores (*LSAT Handbook*, 1964, p. 26; Pitcher, 1965, p. 12). Such a decrease does not, of course, indicate a decline in the validity of the test, but only the increasing utilization of LSAT scores in student selection.

Directions: Each set of data in this section is followed by questions based on its content. After studying a set of data, choose the best answer to each question. . . . Answer each question on the basis of what is *presented* or *implied* in the data which precede it. Use any available space for figuring.

AMOUNT OF HOME BUILDING IN THE UNITED STATES BEFORE 1940



Note: Percentages are based on the total number of homes of a given type built before 1940.
Assume that none of the houses built before 1940 was destroyed.
All reference to homes in this set is limited to homes in the United States before 1940.

1. What fraction of the urban homes were built before 1920?

   (A) $\frac{19}{100}$ (B) $\frac{2}{5}$ (C) $\frac{1}{2}$ (D) $\frac{3}{5}$ (E) $\frac{3}{4}$

2. How many more urban homes than rural nonfarm homes (to the nearest 0.1 million) had been built by the beginning of 1890?

   (A) 1.0 (B) 1.2 (C) 1.4 (D) 2.0
   (E) There was an equal number of each type.

3. At the end of which of the following periods did the total number of urban homes first exceed the total number of rural farm homes?

   (A) Before 1890 (B) 1890-1899 (C) 1900-1909
   (D) 1910-1919 (E) 1920-1929

ANSWERS: 1–D, 2–C, 3–A.

Fig. 86. Typical Data Interpretation Items from Law School Admission Test. (From *Law School Admission Bulletin, 1973–74.* Reprinted by permission of The Law School Admission Council and Educational Testing Service.)

The Writing Ability test, introduced in 1961, adds little to the validity of the battery in predicting first-year law grades (Pitcher, 1965, pp. 15–20). It was included principally on the basis of content validity, in order to select prospective lawyers who are competent in written communication. Like LSAT, it, too, uses multiple-choice items, principally of two types. Error recognition requires the candidate to detect errors in given sentences and to identify the nature of the error (e.g., word choice, grammar, verbosity or redundancy). In sentence correction, the task is to choose the most effective way of phrasing designated portions of given sentences.

CERTIFICATION AND LICENSING OF PSYCHOLOGISTS. The use of tests in assessing the competence of professional personnel after the completion of training can be illustrated within the field of psychology. It has been customary in this field to differentiate between certification, concerned essentially with control of the title "psychologist," and licensing, designed to control the practice of psychology. Nearly all states now have laws of one type or the other. Although the majority began as certification laws, there is a strong movement to change to the more stringent licensing laws, which require a precise definition of the practice of psychology. Clearly the licensing laws are more effective in protecting the public against charlatans and incompetent practitioners. Both types of laws are designed principally for psychologists who render professional services directly to individuals, organizations, or the public rather than as a part of their functions in a university, government agency, or similar institution.

Although varying in details, the basic requirements are similar in most states. We may consider the New York State certification procedures as an example. To be admitted to the psychologist examination, candidates must have a doctoral degree in psychology plus two years of supervised experience. The examination itself has two parts. Part I consists of 100 multiple-choice items distributed over seven areas: clinical, developmental, industrial and personnel, general experimental, measurement and statistics, social and personality, and counseling. Item analyses are regularly conducted in terms of difficulty level, biserial correlation with total score, and percentage of respondents choosing each option. Each form of the examination contains one-half new items and one-half old items selected on the basis of the item-analysis data. Part II comprises five essay questions, one dealing with research methodology and four from any two of the seven areas; these two areas must be chosen by the candidate in advance of the examination. The essay questions, scored independently by two readers, have yielded scorer reliabilities clustering in the .80's.

Most other states administer a uniform, multiple-choice examination developed by Professional Examination Service for the American Association of State Psychology Boards. Different forms of this examination require from four to five hours. The items are designed to assess knowledge of the major fields of psychology at a level that a licensed or certified psychologist should have attained regardless of his own specialty. Other requirements, such as oral examinations, interviews, essay examinations, and qualifications with regard to education and experience, are set by the individual states.

Generally, an individual is licensed or certified for the practice of psychology as a whole, rather than in a specialty. Moreover, these procedures are concerned with the attainment of minimal competence for acceptable practice. A higher level of accreditation is provided by the American Board of Professional Psychology (ABPP), an independently incorporated board within the profession itself, originally established by the American Psychological Association. ABPP grants diplomas in four specialties: Clinical, Counseling, Industrial and Organizational, and School Psychology. Requirements include a PhD in the psychological specialty in which the diploma is sought, plus five years of experience in the specialty. In addition, ABPP now requires an oral examination in the individual's own specialty, which may include field observations of his professional work. Formerly, ABPP also administered a written examination, including both objective and essay parts; but it was discontinued partly because of the increasing diversification of psychological practice, making the preparation of common examinations at such advanced levels very difficult. Nevertheless, the change has aroused some strong criticisms (see, e.g., Newman, 1966).

CHAPTER 16

Clinical Testing

CLINICAL psychologists employ a wide variety of tests, including most of the types already discussed. High in frequency of use are such individual intelligence tests as the Stanford-Binet and the Wechsler scales, as well as some of the nonverbal and performance tests discussed in Chapter 10. Many of the personality tests to be considered in the next four chapters are also prominent in the clinician's repertory. Certain diagnostic educational tests are appropriate in cases of reading disabilities or other learning problems.

Clinicians typically draw upon multiple sources of data in the intensive study of individual cases. Information derived from interviewing and from the case history is combined with test scores to build up an integrated picture of the individual. The clinician thus has available certain safeguards against overgeneralizing from isolated test scores. This fact probably accounts at least in part for the continued use in clinical contexts of some tests whose psychometric properties are either weak or unproved. As long as such instruments serve primarily to suggest leads for the skilled clinician to follow up, their retention can be justified. There is a danger, of course, that a relatively inexperienced and overzealous clinician, unmindful of the limitations of an instrument, may place more confidence in its scores than is warranted.

The effectively functioning clinical psychologist engages in a continuing sequence of hypothesis formation and hypothesis testing about the individual case. Each item of information—whether it be an event recorded in the case history, a comment by the client, or a test score—suggests a hypothesis about the individual that will be either confirmed or refuted as other facts are gathered. Such hypotheses themselves indicate the direction of further lines of inquiry. It should be borne in mind that even highly reliable tests with well-established validity do not yield sufficiently precise results for individual diagnosis. Hence, it is understandable that clinicians as a group tend to be more receptive than other psychologists to psychometrically crude instruments, which may nevertheless provide a rich harvest of leads for further exploration.

In the present chapter, we shall consider a few types of tests used predominantly if not exclusively in clinical settings. For a comprehensive view of clinical testing, these instruments should be added to those mentioned in the opening paragraph. We shall also examine certain special procedures and problems characteristic of clinical testing.

DIAGNOSTIC USE OF INTELLIGENCE TESTS

When an alert and trained clinician is in active contact with an individual during the hour or so required to administer a test, he can hardly fail to learn more about that person than is conveyed by an IQ or some other single score. This is still true if the test itself is administered by a technician, provided that a complete record of the examinee's responses is retained, as is customary in individually administered tests. The literature of clinical psychology contains thousands of research reports and theoretical discussions designed to systematize, facilitate, and improve the utilization of these supplementary sources of data.

PATTERN ANALYSIS OF TEST SCORES. Besides using intelligence tests to assess an individual's general level of intellectual functioning, clinical psychologists also customarily explore the pattern, or profile, of test scores for possible indices of psychopathology. It is likely that pathological deterioration, as from brain damage or psychotic disorders, does not affect all intellectual functions uniformly. Some functions are believed to remain relatively unaffected, while others are considered to be more sensitive to particular pathological influences. Similarly, neurotic anxiety may seriously interfere with performance on certain types of tests, which require careful observation and concentration, while leaving performance on other tests unimpaired.

The Wechsler scales lend themselves especially well to such profile analyses, since all subtest scores are expressed in directly comparable standard scores. From the outset, Wechsler has described a number of diagnostic uses of his scales. Several other clinicians have recommended other techniques and modifications (see, e.g., Guertin et al., 1962, 1966; Matarazzo, 1972; Rapaport, 1945). Specifically, any one of three major procedures have been utilized. The first involves the amount of *scatter,* or extent of variation among the individual's scores on all subtests. The underlying rationale of scatter indices is that intertest variation should be larger in pathological than in normal cases. The second procedure is based on the computation of a *deterioration index* based on the difference between "hold" tests, allegedly resistant to deterioration from pathology or old age, and "don't hold" tests, considered to be susceptible to decline. The third approach is based on *score patterns* associated with particular clinical syndromes, such as brain damage, schizophrenia, anxiety states, and delinquency. Wechsler and other investigators have described patterns of high and low subtest scores characterizing each of these disorders, among others.

Three decades of research on these various forms of pattern analysis with the Wechsler scales have provided little support for their diagnostic value (see Guertin et al., 1956, 1962, 1966, 1971; Matarazzo, 1972). Several methodological requirements must be considered in evaluating this research. First, it is obviously essential to ascertain what is the minimum statistically significant difference between any two test scores. It will be recalled (Ch. 5) that the probability of a difference exceeding chance depends on the reliability coefficients of the two tests. The data required for such statistical evaluation of score differences are now available in the Wechsler scale manuals, as well as in other pertinent publications (e.g., Field, 1960; Zimmerman, Woo-Sam, & Glasser, 1973). For instance, the minimum statistically significant difference between WAIS Verbal and Performance IQ's is 10 points at the .05 level and 13 points at the .01 level. The same limitations apply to comparisons among individual subtest scores, whose reliabilities are generally lower than those of the composite V and P IQ's.

A second consideration is the base rate or frequency of occurrence of any given difference in the standardization sample (Field, 1960). Although a V-P difference of 15 or more IQ points is statistically significant and is generally regarded as clinically suspect, differences of this magnitude occurred in about 10 percent of the standardization sample. A third and related point, pertaining to differences between individual subtests, stems from the large number of possible paired comparisons between tests in any individual's record. As H. G. Jones (1956) observed, the 11 subtests yield 55 possible intertest differences. Hence a difference expected in, let us say, 10 percent of the cases between any one pair of subtests will actually occur about five times in a single person's record (10% of 55 = 5.5). Rather than occurring in only 10 percent of normal persons, therefore, differences of such magnitude would be found, on the average, five times in *every* normal person's record.

A fourth important methodological requirement is the cross validation of group differences in score patterns. Because of the large number of possible diagnostic signs, or pattern differences, the comparison of any single pathological sample with a normal control sample could easily yield a set of spurious differences through sampling error. When actually submitted to cross validation, most of the initial findings in pattern analysis were not, in fact, corroborated.

A fifth problem arises from the multiplicity of conditions that may account for atypical variations among subtest scores. Such variations may result not only from pathology but also from differences in educational, occupational, cultural, or other background factors. Language handicap may account for lower Verbal than Performance scores. It will be recalled that skilled laborers tend to score higher on Performance than on Verbal Scales, unlike white-collar groups (Ch. 9). Socioeconomic and urban-rural differences in subtest pattern have likewise been noted. There is also evidence that education affects relative standing on Verbal and Performance Scales. In a study of psychiatric patients (A. Smith, 1966), as schooling decreased from over 12 years to under 5 years, the percentage of cases with higher Performance than Verbal IQ's increased from 21 to 64. Sex differences and subcultural differences in subtest scores may likewise confound the interpretation of WAIS patterns (Levinson, 1959, 1963; Norman, 1966; Zimmerman, Woo-Sam, & Glasser, 1973).

A sixth point—especially emphasized by Matarazzo (1972)—pertains to the nature of the traditional diagnostic categories that provided the criteria for pattern analysis. In the large majority of studies, these categories were heterogeneous, unreliable, and crude. Such broad designations as brain damage or schizophrenia are probably associated with such varied and contradictory behavioral correlates in different subgroups as to be of little use in the identification of typical test patterns.

PROBLEMS IN THE DIAGNOSIS OF BRAIN DAMAGE. Knowledge about the behavioral effects of brain damage dates largely from the writings of Kurt Goldstein and his associates in the early 1920s. Following extensive observations of soldiers who had sustained brain injuries during World War I, Goldstein formulated his classic description of the intellectual impairment associated with brain damage. Among the principal symptoms were a diminution in the ability for abstract thought and a tendency to respond to extraneous stimuli that may disrupt normal perception. It was on the basis of these observations that Goldstein and his co-workers prepared the series of tests to be described later in this chapter.

Widespread concern with brain injury in children arose in the late 1930s and in the 1940s, following the research of Alfred Strauss and his associates (Werner & Strauss, 1941, 1943; Strauss & Lehtinen, 1947). These investigators identified a subgroup of mentally retarded children whose case histories showed evidence of brain injury due to trauma or infection occurring before, during, or shortly after birth. The behavioral description of these children represented an extension and elaboration of the adult syndrome formulated by Goldstein. It delineated a distinctive pattern of both intellectual and emotional disorders that has been widely accepted as characteristic of *the* brain-injured child. Included in this pattern are specific perceptual and conceptual disorders combined with relatively high verbal ability, as well as overactivity, distractibility, and aggressiveness. For many years, both research and practice with brain-injured children have been dominated by a unidimensional concept of "organicity." This approach led to a search for diagnostic tests of organic involvement as such and an attempt to devise remedial or educational programs suitable for brain-injured children as a whole.

Since midcentury, psychologists have increasingly recognized that brain injury may lead to a wide variety of behavioral patterns. No one symptom or set of symptoms need be common to all brain-injured cases. In fact, brain damage may produce the opposite behavior pattern in two individuals. Such findings are consistent with the wide diversity of the underlying organic pathology itself. A significant advance in the analysis of brain-behavior relationships was made by the research of Reitan and his co-workers at the Indiana University Medical Center (see Matarazzo, 1972, Ch. 13; Reitan, 1955, 1966). These investigations provide a good example of the importance of working with relatively homogeneous and clearly defined criterion groups.

In an initial attempt to sort out the variety of disorders loosely classified under "brain damage," Reitan selected for study three groups of patients with left-hemisphere lesions, right-hemisphere lesions, and diffuse brain damage involving both hemispheres, respectively. Diagnoses were based on neurological and neurosurgical data, utilizing electroencephalography and other sophisticated techniques. The three groups were approximately comparable in age and education. Although the samples were small, the principal findings have been corroborated in a number of subsequent investigations (Guertin et al., 1962, 1966; Matarazzo, 1972, p. 390). Essentially, the results showed that left-hemisphere lesions tend to be associated with lower Verbal than Performance IQ on the Wechsler scales ($V < P$). The opposite pattern ($V > P$) predominated in the groups with lesions in the right hemisphere and with diffuse brain damage. The latter score pattern corresponds to the classic diagnostic pattern of brain damage.

Research by Reitan and other investigators is continuing to sort out the complex interactions of other variables with the behavioral effects of brain pathology. There is a growing body of evidence indicating that *age* affects the behavioral symptoms resulting from brain damage. It cannot be assumed that brain damage will have the same effect in an adult, a school-age child, and an infant. Since the nature of intelligence varies at different age levels, the pattern of impairment may vary as a function of the age at which the injury is sustained. The behavioral effects will also depend upon the amount of learning and intellectual development that has occurred prior to the injury. Research on preschool children, for example, indicates that at this age level the brain-injured tend to be deficient in *all* intellectual functions (Graham, Ernhart, et al., 1963). Unlike the pattern found at older ages, impairment among brain-injured preschoolers was just as great in vocabulary as in other cognitive and perceptual functions. Similarly, other studies of young children with a history of prenatal or birth trauma have found an average reduction in IQ on the Stanford-Binet and other standardized intelligence tests. Such broad intellectual impairment may occur because the brain damage itself is diffuse or because a critical deficiency in linguistic development or attention control may seriously hinder the acquisition of other abilities. It is also interesting to note that studies with brain-injured preschool children have failed to find the emotional symptoms of the classic Strauss pattern (Graham et al., 1962; Graham, Ernhart, et al., 1963).

Chronicity has also been found to affect test performance. Available data suggest that the relationships found by Reitan and others may hold for acute but not for chronic cases, that is, for recent rather than for long-standing lesions. The amount of time that has elapsed since the injury may be related not only to progressive physiological changes but also to the extent of behavioral recovery through learning or compensatory readjustments.

Finally, it should be noted that in some cases intellectual impairment may be an *indirect result* of brain damage. Throughout the individual's development, organic and experiential factors interact. Some of the personality disorders included in the classic picture of brain-injured children, for example, may be an indirect effect of the frustrations and interpersonal difficulties experienced by the child with an organically caused intellectual deficiency. Whether or not these personality disorders develop may thus depend on the degree of understanding and the attitudes exhibited by parents, teachers, and other significant persons in the child's environment.

Another example of indirect effects of organic impairment is provided by the intellectual retardation often found among children with cerebral palsy (see, e.g., Sarason & Gladwin, 1959, Ch. 7). In some of these cases, the lesion extends to the cortical level. Both motor and intellectual disorders in these cases result directly from organic damage. In other cases, however, the lesion may be limited to subcortical levels, directly causing only the motor handicaps. If these motor handicaps are severe enough, they may seriously interfere with the development of speech and writing, as well as locomotion and other gross motor activities. In these cases, intellectual retardation is likely to result from educational and social handicaps. Through special educational procedures that bypass the motor handicaps, however, these children can reach a normal or even superior intellectual level, as evidenced by the impressive attainment of some cerebral-palsied individuals.

It is abundantly evident that brain damage covers a wide variety of organic disorders, with correspondingly diverse behavioral manifestations. The test performance of brain-injured persons can be expected to vary with the source, extent, and locus of the cerebral damage; the age at which the damage occurred; the age when the individual's behavior is being assessed; and the duration of the pathological condition. To expect behavioral homogeneity among the brain-injured would thus be highly unrealistic.

From another angle, the same intellectual or other behavioral disorder —and the same diagnostic sign in test performance—may result from organic factors in one person and from experiential factors in another. These experiential factors, moreover, may be unrelated to the organic damage in one case and an indirect result of it in another. It follows that to interpret any specific diagnostic sign in test performance requires additional information about the individual's experiential background and his personal history.

QUALITATIVE OBSERVATIONS. Any individual intelligence test can provide information at various levels. At the most objective level, such scales yield an IQ with high reliability and fair evidence of validity for many purposes. At a less objective level, we find the innumerable attempts at semiquantitative pattern analysis. Both methodological evaluations and empirical results have thus far lent little or no support to the proposed interpretations. Even if well-designed future investigations should conclusively demonstrate a relationship between test response patterns and clearly defined diagnostic categories, the results would still indicate only group trends and would not be equally applicable to all individual cases. Such diagnostic signs, therefore, could not be employed routinely, by formula. They would still need to be interpreted in the light of other data about the individual.

At a purely qualitative level, any irregularity of performance should suggest avenues for further exploration. Applications of this approach to the Wechsler scales can be found in Glasser and Zimmerman (1967), Matarazzo (1972, Ch. 15), and Zimmerman, Woo-Sam, and Glasser (1973.) For each subtest, these sources provide an interpretive framework, with summaries of relevant data and a discussion of the types of information a clinician may obtain with regard to both intellectual and personality characteristics. Similar guidelines have been prepared for the qualitative interpretation of the Stanford-Binet (see, e.g., Sattler, 1974, Ch. 19).

Significant leads may emerge from the form as well as the content of test responses. Bizarreness, overelaboration, or excessive self-reference, for example, may indicate personality disorders. A qualitative analysis of both errors and correct responses may provide useful cues about problem-solving approaches, conceptual development, or cognitive styles. Atypical content of test responses is a further source of leads. As Wechsler (1958, p. 181) observes, if in the Vocabulary test one person defines "sentence" as a group of words and another as a penalty imposed by a judge, this difference may furnish a clue to important dissimilarities in experiential background or personality. Sattler cites an example of a sexually delinquent girl who, when given the Word Naming test of the Stanford-Binet, responded with 35 boys' names (Lejeune, cited in Sattler, 1974, p. 264). Still another source of qualitative data available during the administration of an individual intelligence test is the individual's general behavior in the testing situation. Examples include motor activities, speech, emotional responses, attitude toward examiner, and behavior toward the test materials and the testing environment.

Because of their idiosyncratic nature, such qualitative cues cannot be validated by quantitative methods adapted to the measurement of group trends. The validity of the inferences they suggest can only be tested for the individual case by the process of successive data gathering and verification characteristic of the clinical method.

SPECIAL TESTS FOR DETECTING COGNITIVE DYSFUNCTION

A number of tests have been specially designed as clinical instruments for assessing intellectual impairment. Several have been described as indices of organicity, or brain damage. Most available tests in this category, however, are broader in function, and have been employed to detect intellectual deterioration or impairment arising from a variety of possible causes. Some are designed only to afford the clinician an opportunity for qualitative observations. Others have relatively standardized procedure and objective norms, although the normative samples are often inadequate. Some are subject to one or more of the theoretical objections raised against profile analysis of Wechsler scores. A few have yielded promising empirical data.

All available tests of intellectual impairment are based on the premise of a differential deficit in different functions. Chief among the functions considered to be most sensitive to pathological processes are memory, spatial perception, and abstraction or concept formation. Three typical tests will be examined in this section.

BENTON VISUAL RETENTION TEST. This test utilizes 10 cards, each containing one or more simple geometric figures. In the standard administration, each card is exposed for 10 seconds and the examinee is told to draw what was on the card immediately after its removal. The test thus requires spatial perception, immediate recall, and visuomotor reproduction of drawings. Performance is scored in terms of number of cards correctly reproduced and total number of errors. Interscorer reliability of the order of .95 is reported for these scores. Additional qualitative information can be obtained by classifying the errors as omissions (and additions), distortions, perseverations, rotations, misplacements, and size errors (see Fig. 87). Three equivalent forms of 10 drawings each are available. Reliability assessed by correlation of alternate forms is in the .80's.

The Benton test may also be administered with certain procedural variations. Of particular interest is the copying administration, in which the card remains in view while the examinee executes his drawing. This procedure thus permits a separation of perceptual from memory errors. In the interpretation of scores, both number of correct reproductions and number of errors may be compared with the expected "normal" score for each age and intellectual level. The latter can be determined from the IQ on any standard verbal intelligence test. Benton scores falling more than a designated number of points below the expected level for an individual are considered significant in the clinical sense. The manual makes it clear,

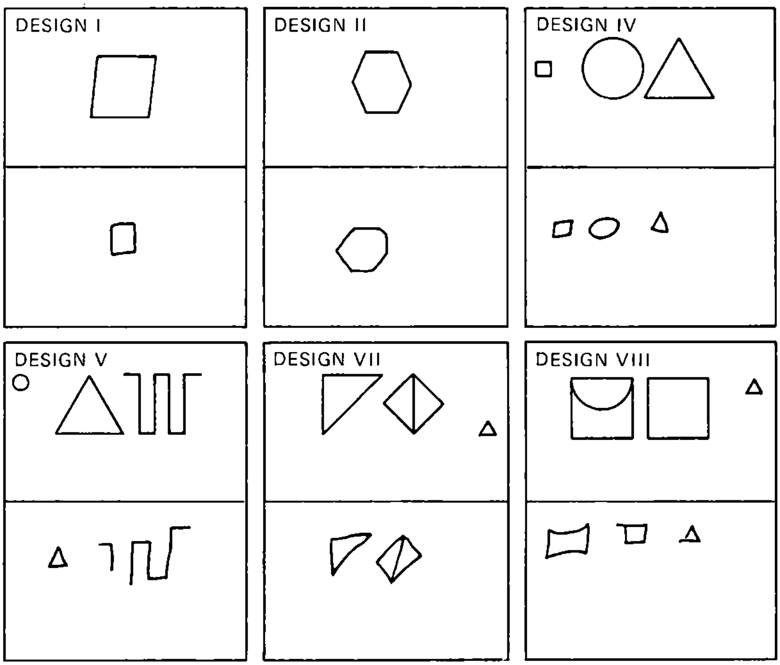


Fig. 87. Errors by a Brain-Injured Patient on Benton Visual Retention Test (Copying Administration). Upper design in each item is sample, lower is patient's copy. In addition to distortions and size errors, note omission of small peripheral figures in Designs V and VII.

(From Benton, 1974, p. 63. Reproduced by permission. Copyright 1955, 1963, © 1974 by The Psychological Corporation, New York, N.Y. All rights reserved.)

however, that many conditions other than pathology could account for a deviant performance on this test. To reach a diagnosis, the clinician needs corroborative data from other appropriate tests, together with information about the individual's history and background.

Several investigations have yielded significant mean differences between brain-injured and control cases in both number of correct drawings and total number of errors. While overlapping in score distributions is always present, the frequency of correct differentiations appears promising. In one comparison of 100 brain-injured and 100 control cases, for example, a cutoff score of 3 points below expected number of correct drawings identified 57 percent of the patients, with only 4 percent false positives among the controls. When the cutoff was raised to 4 points below the expected score, 36 percent of the patients and none of the controls were identified as brain-injured (Benton, 1974, pp. 52–53). Although useful in detecting brain-injury in children, the test does not differentiate as sharply when applied to children as when applied to adults. Suggestive data have also been reported on the performance of several other groups, such as schizophrenics, emotionally disturbed children, mental retardates, and persons over 65.

BENDER-GESTALT TEST. The Bender Visual Motor Gestalt Test, commonly known as the Bender-Gestalt Test, is widely used by clinical psychologists, predominantly for the detection of brain damage. In this test, the nine simple designs shown in Figure 88 are presented one at a time on cards. The examinee is instructed to copy each design, with the sample before him. The designs were selected by Bender from a longer series originally employed by Wertheimer, one of the founders of the Gestalt school, in his studies of visual perception. The particular designs were constructed so as to illustrate certain principles of Gestalt psychology, and Bender's own analyses of the test results are formulated in terms of Gestalt concepts. Although for many years the test was administered by Bender and others to children and adults showing a variety of intellectual and emotional disorders, the data were not reported in objective and systematic form and were therefore difficult to evaluate.

While many clinicians still interpret the Bender-Gestalt Test through subjective intuitive procedures, several objective scoring systems have been developed for use with either adults or children (see Tolor & Schulberg, 1963). In one of the most carefully developed of these scoring systems, Pascal and Suttell (1951) undertook a standardization and quantification of the Bender-Gestalt Test on an *adult population*. On the basis of the drawing errors that significantly differentiated between matched samples of normals and abnormals, a relatively objective scoring key was developed. Cross validation of this key on new samples of 474 nonpatients (or normal controls), 187 neurotics, and 136 psychotics yielded mean scores of 50, 68.2, and 81.8 respectively. These scores are standard scores with a mean of 50 and an *SD* of 10, the higher scores indicating more diagnostic errors. The biserial correlation of test scores against the criterion of patient versus nonpatient status was .74. This correlation may be regarded as a measure of concurrent criterion-related validity. There is also evidence that this adaptation of the test can significantly differentiate groups of organics from both normal and psychotic groups (Tolor & Schulberg, 1963).

Retest reliabilities of about .70 were found in normal samples over a 24-hour interval. Scorer reliabilities of approximately .90 are reported for trained scorers. Performance on the test is apparently independent of drawing ability, but is significantly related to amount of education. In this adaptation, the Bender-Gestalt Test appears to have promise as a rapid screening device, especially for detecting the more serious forms

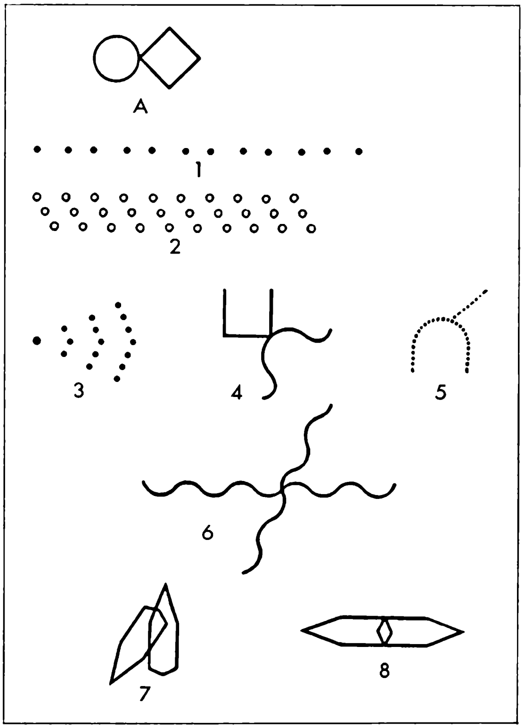


Fig. 88. The Bender-Gestalt Test.
(From Bender, 1938, p. 4; reproduced by permission of Lauretta Bender.)

of disturbance. The normative sample, however, is rather restricted geographically, educationally, and in other ways. Extension and revision of the norms on the basis of a larger and more representative sample would be desirable. Continuing research on validity in different samples and against other criteria is also needed.

Koppitz (1975) carried out an extensive standardization of the Bender-Gestalt Test with *children.*[1] Providing norms on 1,104 kindergarten to fourth-grade children in midwestern and eastern public schools, this adaptation of the Bender-Gestalt Test was prepared as a nonverbal developmental scale for ages 5 to 10. Scorer reliability appears satisfactory, the interscorer correlations varying between .88 and .96. Retest reliability over a 4-month interval within single grade groups was rather low, with coefficients ranging from .547 to .659. Between the ages of 5 and 10 years, the scores show consistent improvement with age and moderate to high correlations with standard intelligence tests. Within single-year groups, correlations ranging from .48 to .79 were found between Bender-Gestalt scores and Stanford-Binet or WISC IQ's. After the age of 10, the Bender-Gestalt no longer correlates significantly with either intelligence test scores or age, since normal persons beyond that age obtain virtually perfect scores.

Koppitz reports fairly high validities for the test in assessing school readiness and in predicting the subsequent educational achievement of first-grade children. Studies of first- and second-grade children also showed significant relationships between Bender-Gestalt scores and performance in reading and arithmetic. Finally, evidence is given suggesting that among mentally retarded children the Bender-Gestalt developmental score has fairly high validity as a measure of intellectual level and as a predictor of academic achievement. All these analyses are based on small samples and need further verification.

Comparison of brain-injured with normal children between the ages of 5 and 10 showed significant group differences in total score on the Bender-Gestalt developmental scale. For a diagnosis of brain injury, however, Koppitz concluded that the total score should be supplemented with a number of additional observations of the child's performance, including time required, amount of space employed in reproducing the drawings, detailed analysis of individual errors, observations of the child's behavior, and inquiry into the child's awareness of his errors. All the uses of the Koppitz adaptation of the Bender-Gestalt discussed thus far are based on the developmental scoring scale. In addition, Koppitz presents a set of 10 "emotional indicators" applicable when the test is employed as a projective instrument for detecting emotional disturbances in children.

---

[1] Also known as The Bender Gestalt Test for Young Children.

GOLDSTEIN-SCHEERER TESTS OF ABSTRACT AND CONCRETE THINKING. Clinical tests of conceptual thinking are more concerned with the methods employed by the respondent than with the end result achieved. Largely for this reason, such tests lean heavily on qualitative observations and have thus far made little use of objective scoring and standardized norms. It is recognized, of course, that normative data would be helpful even in the interpretation of qualitative observations, but such data are difficult to obtain.

Developed by Goldstein and his associates as a result of their experience with brain-damaged patients during and after World War I, the Goldstein-Scheerer series still enjoys wide popularity in clinical practice. Goldstein and his co-workers concluded that the principal effect of brain damage is an impairment of the "abstract attitude." This impairment, they maintained, could be manifested in a number of ways, such as the inability to abstract common properties from objects, to grasp the essential quality of a given whole, to break up a given whole into parts, or to shift from one aspect of a situation to another. The series consists of five tests, illustrated in Figure 89 and described below:

1. *Goldstein-Scheerer Cube Test.* This is an adaptation of the Kohs Block Design Test, which was also the basis of the Block Design subtest of the Wechsler scales. Twelve designs are to be copied, all but one being taken from the original Kohs series. The designs are presented in increasing order of difficulty. If the examinee fails to copy a design, he is given a series of aids which render the task progressively easier and less abstract. For example, lines are introduced on the model design to show where the blocks fit together; or a larger model design is substituted that corresponds to the size of the block design; or an exact replica in blocks is substituted for the printed model. Ability to transfer what has been learned by means of these aids to the performance of the original task is also tested.

2. *Weigl-Goldstein-Scheerer Color-Form Sorting Test.* The materials for this test consist of 12 pieces, including 4 circles, 4 squares, and 4 triangles, each shape occurring in red, green, yellow, and blue. The 12 pieces are presented in random order, with the instructions to place together those that belong together. Upon completion of this task, the examinee is asked to sort the pieces in another way. If he does not spontaneously sort the pieces according to form and color, a number of aids are provided to facilitate the task, as in the Cube test. In this test, interest is focused on the person's ability to shift from one basis of classification to the other.

3. *Goldstein-Scheerer Stick Test.* Thirty sticks of 4 lengths are used in this test. The examinee is asked to copy simple geometric designs composed of sticks, and is subsequently required to reproduce the same figures

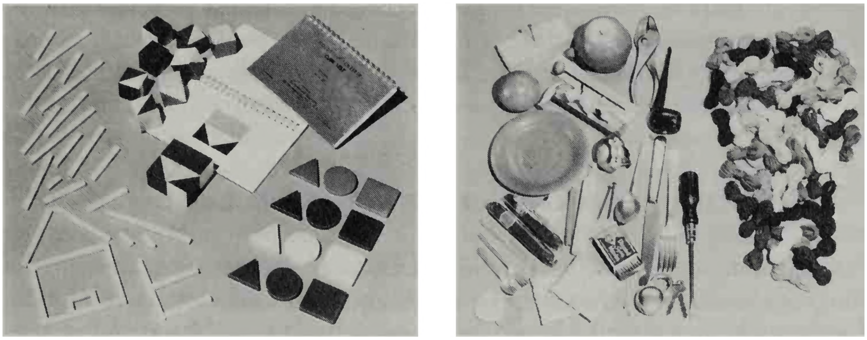


Fig. 89. Goldstein-Scheerer Tests of Abstract and Concrete Thinking. From left to right: Stick, Cube, Color-Form Sorting, Object Sorting, and Color Sorting tests.

(Courtesy The Psychological Corporation.)

from memory. The designs are presented in increasing order of difficulty. Even more than in the other tests in the series, the evaluation of performance on this test depends on the examinee's oral explanation of his procedure.

4. *Gelb-Goldstein Color Sorting Test.* The examinee is shown a collection of colored woolen skeins, in several different shades of each color. In successive parts of the test, he is instructed to select all the others that go with the given sample or samples. Either color or brightness may serve as the basis for classification at the abstract level; a more concrete type of response is characterized by a simple matching procedure.
5. *Goldstein-Scheerer Object Sorting Test.* Figure 89 shows the objects employed in this test. Several different sorting situations are presented in its administration. For example, the examinee is given an article selected by the examiner and is asked to pick out the others that belong with it. Or he is asked to place all the articles into groups of his own choice. The examiner also places objects into groups on some basis, such as color, form, use, etc., and asks the examinee to state the basis of the classification.

While describing fairly standardized procedures for administering these tests, the authors have developed no standardized scoring system. They give general criteria for rating various levels of abstraction in the examinee's responses to each test. In addition, they list a number of specific anomalies of test performance, but empirical validation of the diagnostic significance of these anomalies is meager. For some of the tests in the series, other investigators have devised objective scoring systems, chiefly on the basis of research with schizophrenics. The Goldstein-Scheerer tests have been found to correlate highly with scores on standard verbal intelligence tests in normal persons. For an adequate assessment of intellectual impairment, therefore, performance on the Goldstein-Scheerer tests should be evaluated against the individual's earlier intellectual level or on some independent index of verbal intelligence.

IDENTIFYING SPECIFIC LEARNING DISABILITIES

THE PROBLEM. The 1970s ushered in a wave of crash programs for the diagnosis and remediation of learning disabilities. Educators have become increasingly aware of the high frequency of this type of handicap among schoolchildren. In the legislation of the Bureau of Education for the Handicapped, these children are described as having,

> . . . a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations. Such disorders include such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. Such term does not include children who have learning problems which are primarily the result of visual, hearing, or motor handicaps, or mental retardation, or emotional disturbance, or environmental disadvantage.[2]

It follows from this definition that the designation of learning disability should not be employed until several other conditions have been eliminated as possible causes of the child's educational or psychological difficulties. The varied terminology applied to these cases reflects both a changing approach over time and a convergence of medical, educational, and psychological orientations. What are now characterized as learning disabilities were until recently regarded as indices of brain injury or, at least, of minimal brain dysfunction. The current view is that neurological pathology *may* underlie learning disabilities; but its presence cannot be assumed in all cases. Moreover, learning disability specialists characteristically insist that they have no interest in etiology but want only to pinpoint the specific pattern of behavioral disabilities in each case in order to plan an appropriate program of remedial instruction. The teacher, for example, needs to know the precise locus and limits of a child's perceptual disabilities so that she may "teach to strength" by utilizing the child's strongest channel, or sense modality, at the outset.

Nevertheless, more knowledge about the underlying causes of learning disabilities should contribute to the accuracy of diagnosis and enhance the effectiveness of the treatment program in individual cases. Partly for this reason, available procedures for both diagnosis and remediation of learning disabilities need to remain flexible and responsive to new advances. It should also be recognized that there are still wide individual differences in theoretical orientation among specialists on learning disabilities. These differences are reflected in both testing instruments and remedial programs.

---

[2] Cited by Swets and Elliott (1974, pp. 1–2), from Public Law 91–230 (89 Stat. 177), Section 602–15, April 13, 1970.

Typically, children with learning disabilities show normal or above-normal intelligence, in combination with pronounced difficulties in learning one or more basic educational skills (most often, reading). In addition, they manifest various combinations of associated behavioral symptoms. Chief among such symptoms are perceptual disorders in one or more sense modalities, as well as poor integration of input from different modalities and disruption of sensorimotor coordination. These perceptual disorders are often directly related to their reading difficulty and other learning problems. Deficiencies in memory, attention control, and conceptual skills are common.

Many disturbances of language development may occur. Aphasia originally referred to a loss in the ability to understand language (receptive aphasia) or to use language meaningfully (expressive aphasia) following a brain injury. In the case of a young child, however, these abilities may not have been acquired in the first place—hence the disorder is designated as "developmental aphasia." Other typical symptoms include various forms and degrees of motor incoordination (affecting gross or fine motor control or both); temporal and spatial disorientation; difficulty in organizing activities and following a plan; and undirected, impulsive hyperactivity. Aggression and other emotional and interpersonal problems may also develop, often as a direct result of the academic failures and frustrations engendered by the child's learning disabilities. In evaluating the child's behavior, it should also be borne in mind that many specific difficulties that are normal at an early age (e.g., in a three-year-old) represent a dysfunction if they persist to an older age. A developmental frame of reference is therefore needed, with qualitative if not quantitative norms.[3]

ASSESSMENT TECHNIQUES. Regardless of theoretical orientation, there is general agreement that the identification of learning disabilities requires a wide assortment of tests and supplementary observational procedures. This follows from at least three features of the diagnostic problem: (1)

---

[3] This description is necessarily limited. For a fuller discussion, see any of the many recent books on learning disabilities, such as Johnson and Myklebust (1967), Kirk (1972), Lerner (1971), McCarthy and McCarthy (1969), Waugh and Bush (1971), and Wender (1971).

the variety of behavioral disorders associated with this condition; (2) individual differences in the particular combination of symptoms; and (3) the need for highly specific information regarding the nature and extent of the disabilities in each case.

Usually, the assessment of children with learning disabilities is a cooperative effort by a professional team. The classroom teacher can administer group tests and employ other screening or wide-band instruments. Examples of screening techniques include Slingerland's Screening Test for Identifying Children with Specific Language Disability, an eight-test group battery yielding 21 scores; and Myklebust's Pupil Rating Scale, a behavioral rating scale based on day-by-day classroom observations. Wide-range achievement tests can also be administered by teachers, either individually (e.g., Peabody Individual Achievement Test) or to groups (e.g., Jastak's Wide Range Achievement Test).

Regular achievement batteries may also serve in this connection, particularly those designed for primary and preschool levels and permitting criterion-referenced analyses of specific strengths and weaknesses. Several instruments cited in Chapter 14 are appropriate for this purpose, including the primary levels of the Stanford Achievement Test, the Stanford Early School Achievement Test (SESAT), the Metropolitan Readiness Tests, and Circus. An examination of all such instruments discloses many subtests and item types that closely resemble those found in tests specially designed for identifying learning disabilities. For older children, the various diagnostic tests in reading and arithmetic provide highly relevant information (see Ch. 14 and Appendix E).

The administration and interpretation of individual intelligence tests require the services of a psychologist. Such tests as the Stanford-Binet, WISC, WPPSI, and McCarthy Scales of Children's Abilities not only provide a global index such as an IQ to help differentiate between mental retardation and special learning disabilities, but they also yield much qualitative information regarding specific deficiencies (see, e.g., Kaufman & Kaufman, 1974; Sattler, 1974, Chs. 17, 21, 23; Waugh & Bush, 1971). For example, these tests can reveal possible deficiencies in the perception and recall of visual patterns, motor difficulties in the copying of forms, limitations of short-term memory, inability to handle abstract concepts, and many types of language disorders.

The intensive study of each individual case in order to map out the specific pattern of learning disabilities is generally the responsibility of either an educational specialist on learning disabilities or a psychologist. The batteries assembled for this purpose usually include some test of aphasia. Other tests covering the many aspects of the understanding and use of written and spoken language are of particular importance. Tests designed to detect disturbances of perception and short-term memory, such as the Bender-Gestalt and the Benton Visual Retention Test, are highly appropriate, as are tests of auditory discrimination. Motor functions also need to be covered (see, e.g., Lincoln-Oseretsky, Ch. 10).

In addition to these standard instruments, a number of new tests have been prepared specially for children with learning disabilities (see Appendix E).[4] All these tests should be regarded, not as psychometric instruments, but as observational aids for the clinical psychologist and learning disabilities specialist. Most provide no norms. When normative data are available, they are usually derived from small and narrowly limited samples. Moreover, such norms serve principally as reference points for defining "normal" (i.e., nonpathological) responses rather than for the quantitative evaluation of performance. Although not explicitly labeled as such, most of these tests are criterion-referenced insofar as they provide a description of the child's deficiencies in highly specific terms. Primary focus is on whether the child has mastered the skills displayed by normal children of his age.

One of the most carefully constructed tests in this area, on which a relatively large amount of research has already been conducted, is the Illinois Test of Psycholinguistic Abilities (ITPA). This is an individual test for children between the ages of 2 and 10 years. In its design, the ITPA follows the three-dimensional model shown below, which is a condensed adaptation of C. E. Osgood's theoretical model of the communication process (Kirk & Kirk, 1971):

Channels:	auditory-vocal
	visual-motor
Processes:	receptive (understanding words and pictures)
	organizing (association of past and present inputs and eduction of relations)
	expressive (expressing ideas in words or gestures)
Levels:	representational (utilizing meaning of linguistic symbols)
	automatic (utilizing habitual integrated patterns—e.g., speed of perception, rote learning, memory span)

The abilities covered by the ITPA are located at the intersections of the three dimensions. For example, in the Manual Expression Test (visual-motor channel, expressive process, representational level), the child performs manual gestures to "show what we do with" each pictured object, such as pencil sharpener, telephone, and clarinet. In the Grammatic Closure Test (auditory-vocal channel, organizing process, automatic level), the task is to complete oral statements such as "Here is a child. Here are three ———," while shown the appropriate pictures. The entire battery consists of 10 regular subtests, plus two optional subtests.

---

[4] In the *Mental Measurements Yearbook*, such tests are usually listed under "Sensory and Motor Tests" and "Speech and Hearing Tests." Many can be found in the catalogues of major test publishers, especially Consulting Psychologists Press and Western Psychological Services, which include also some tests published by others.

Both the pictorial and verbal items of the ITPA appear to be heavily loaded with culturally restricted content. The normative samples were drawn largely from middle-class populations, in five midwestern cities. For these reasons, the applicability of the ITPA to children from lower socioeconomic levels and from minority groups needs to be more fully ascertained. The test authors recommend the use of supplementary local norms.

Other batteries that provide somewhat different emphases and combinations of functions are the Frostig Developmental Test of Visual Perception[5] and Ayres' Southern California Sensory Integration Tests. The former comprises five tests of visuomotor coordination, visual perception, and spatial relations; the latter is a 17-test battery covering sensorimotor and perceptual tasks, with considerable involvement of both fine and gross motor coordination. Still a different approach is illustrated by Valett's Psychoeducational Inventory of Basic Learning Abilities. Utilizing simple materials and procedures suitable for use by classroom teachers, this instrument provides a rough evaluation of performance in 53 skills grouped into six major areas: gross motor development, sensorimotor integration, perceptual-motor skills, language development, conceptual skills, and social skills. All three batteries form part of comprehensive diagnostic-and-remediation or testing-and-teaching programs developed by their authors.[6]

CLINICAL JUDGMENT

THE NATURE OF CLINICAL ASSESSMENT. The term "clinical" is conventionally employed to refer to any methodology involving the intensive study of individual cases. Thus clinical procedures are applied, not only by psychologists working in clinics, but also by those who function in counseling centers, schools, and industry (e.g., in the assessment of personnel for high-level jobs). A distinguishing feature of clinical assessments is their reliance on *judgment* in at least some aspects of the process. In this respect, the clinical method has been contrasted with the administration of objective, standardized tests and the use of statistical, or actuarial procedures for combining data through regression equations, multiple cutoffs, and other "mechanical formulas" (Ch. 7).

---

[5] Also available in an experimental edition: Frostig Movement Skills Test Battery.

[6] Testing and training materials for these programs are listed in the catalogues of Consulting Psychologists Press (Frostig, Valett) and Western Psychological Services (Ayres).

What the clinician does in assessing a client may be regarded as a special case of person cognition, or interpersonal perception—the process through which anyone comes to know and understand another person (Bakan, 1956; Gage & Cronbach, 1955; Sarbin, Taft, & Bailey, 1960). In this process, the observer often relies upon *assumed similarity* to himself. Thus he can utilize his own experience in interpreting the behavior of another. Although he cannot directly observe another's aches and pains or feelings of joy and sadness, he can identify them through facial expressions, gestures, verbal reports, and other overt cues that he has learned to associate with his own feelings. Assumed similarity between a newcomer and familiar others, such as one's relatives, friends, or former clients, is an extention of this mechanism. When this approach is followed in trying to understand someone quite unlike oneself—or unlike one's earlier acquaintances—it is likely to prove misleading. Errors may thus arise when clinicians make diagnostic or prognostic inferences about a client whose cultural background, education, or socioeconomic level differs markedly from their own (Wainwright, 1958).

THE CLINICIAN'S DATA-GATHERING FUNCTION. The diagnostic role of the clinician can be described in terms of data gathering and data synthesis. We shall begin by considering what the clinician can contribute to the process of data gathering. First, by establishing and maintaining rapport, he may elicit from the client pertinent facts about his life history not readily accessible in other ways. Such life-history data provide a particularly sound basis for understanding an individual and predicting his subsequent behavior (Dailey, 1960). The life history has been aptly described as "an unbiased population of events which is as convincing an operational definition of a person as one can hope for" (Dailey, 1960, p. 21). It might be added that the more data one has from the life history, the less filling-in needs to be done in the interpretive and synthesizing process.

In his search for facts, the clinician is guided partly by the client's own responses, which provide clues for further exploration. Particular responses, when considered in the light of personality theory, relevant research findings, and professional experience, lead to the formulation of tentative hypotheses. These hypotheses are then tested by gathering or examining additional pertinent data. As a result, they may be strengthened, modified, or discarded. Such interlocking of hypothesis formation and hypothesis testing represents both an advantage and a potential hazard of clinical assessment. On the one hand, it permits more flexibility of search and more effective utilization of cues than would be possible with a test, questionnaire, or other standardized procedure. On the other hand, if the clinician is unduly influenced by his early hypotheses, he may look only for data that support them. By the type of questions he asks and the way he formulates them or by subtle expressions of agreement or disagreement, he may influence what the client reports. Sarbin and his associates (1960) call this process "soliciting" as contrasted to "probing." Such biased data-gathering techniques probably account for the remarkably uniform etiologies found among the clients of some psychoanalysts.

A third way in which the clinician contributes to the fact-finding process is by serving as a stimulus in an interpersonal situation. In this regard, the clinical interview functions as a situational test or simulation. It provides a sample of the client's interpersonal behavior, observed under more or less controlled conditions.

THE CLINICIAN'S DATA-SYNTHESIS FUNCTION. Much has been written about the data-processing, synthesizing, or interpretive role of the clinician. Some have proposed a special process of "clinical intuition" that is qualitatively different from other forms of inference. Belief in such a process arises partly from the fact that clinicians are often unable to report the cues they employ in reaching a conclusion (Sarbin et al., 1960, Ch. 8). The mystery disappears, however, when we realize that many of our perceptions utilize cues that are inaccessible to self-examination. A classical example is visual depth perception. We regularly perceive the world in three dimensions without being able to specify the cues we employ in the process. Similarly, after exposure to a test protocol, a set of test scores, a case history, or a face-to-face interaction with a client, the clinician may assert that the patient is creative, or a likely suicide, or a poor psychotherapy risk, even though he cannot verbalize the facts he used in reaching such a conclusion. Being unaware of the cues that mediated his inference, the clinician may also be unaware of the probabilistic nature of the inference and may feel more confidence in it than is justified.

Research on the process of clinical judgment has thrown considerable light on the relation between the data input (test scores, specific responses, score patterns, case history items, etc.) and the clinician's output (judgment, diagnosis, prediction). Correlation and regression analyses can indicate the relative contribution of different inputs to the final judgment; and they can show whether the contribution is linear (additive) or is best represented by some nonlinear combination, as in the use of pattern or configural analysis, or moderator variables (Hammond, Hursch, & Todd, 1964; J. S. Wiggins, 1973, Ch. 4; N. Wiggins & Hoffman, 1968). The results show that a linear regression equation (simple additive model) can closely reproduce the clinician's output in most cases. When a more complex model fits the data better, the differences are so slight as to be negligible (Dawes & Corrigan, 1974; Goldberg, 1968b). Clinicians' own reports of the relative weights they assign to different inputs and the ways in which they combine data may differ sharply from their actual use of the available information. Other research has illuminated some of the possible sources of error in clinical judgment, such as the influence of cultural stereotypes (Chapman and Chapman, 1967; Golding & Rorer, 1972) and reliance on fallacious prediction principles (Kahneman & Tversky, 1973). Examples of the latter include failure to consider base rates or regression effects and assumption that more highly intercorrelated predictors yield higher validity.

COMPARATIVE STUDIES OF CLINICAL AND STATISTICAL PREDICTIONS. Given the same set of facts, such as test scores or life-history data, will clinical judgment provide more accurate predictions of subsequent behavior than would be obtained by the routine application of a regression equation or other empirically derived formula? The question has practical as well as theoretical importance because, once a regression equation or other statistical strategy has been developed, it can be applied by a clerk or a computer, thus freeing the clinician for therapy, research, and other functions. In a provocative book entitled *Clinical Versus Statistical Prediction*, Meehl (1954) discussed the process of clinical judgment and surveyed some 20 investigations comparing the two types of prediction. The criteria predicted in these studies included principally success in some kind of schooling or training (college, Air Force pilot training, etc.), response to therapy on the part of psychotic or neurotic patients, and criminal recidivism as well as institutional adjustment of reformatory inmates. Predictions were made by clinical psychologists, counselors, psychiatrists, and other professional persons with varying amounts of experience in the use of clinical procedures. Focusing only on the process of combining data, rather than on differences in the kind of data obtained, Meehl showed that, with only one questionable exception, the routine application of statistical procedures yielded at least as many correct predictions as clinical analysis, and frequently more.

The publication of Meehl's book stimulated a lively controversy that is still active. It was also followed by further research using better controlled procedures and more clinically relevant criteria (see Sawyer, 1966; J. S. Wiggins, 1973, Ch. 5). A later review of available research led Meehl (1965) to reaffirm his earlier conclusion. Of some 50 studies examined up to that time, about two thirds favored statistical prediction and the remaining third showed no significant difference.[7] In a particularly well-designed study, Goldberg (1965) compared clinician's predictions from the score profile on the Minnesota Multiphasic Personality Inventory (see Ch. 17) with predictions made from a series of statistical formulas. Comparisons of validity coefficients showed that the statistical formulas clearly outperformed the clinicians. The highest validity coefficient in the study was obtained with an empirically derived, linear combination of scores. Another interesting finding, known as "bootstrapping," is that an actuarial model of a clinician's own behavior will usually yield more valid predictions than does the clinician himself (Dawes & Corrigan, 1974; Goldberg, 1970). Reflecting principally the regularities in the clinician's behavior, the model reduces the influence of random procedural errors.

It is important to bear in mind that all these studies dealt only with methods of synthesizing data, when clinical and statistical procedures are applied *to the same data.* A major contribution of the clinical method, however, is the obtaining of data in areas in which satisfactory tests are unavailable. Systematic interviewing, case histories, and direct observation of behavior are still the principal sources of information on many aspects of personality. None of the studies on clinical versus statistical prediction reflect on the clinician's data-gathering function.

In the second place, the clinical method is better suited than the statistical method to the processing of rare and idiosyncratic events whose frequency is too low to permit the development of statistical strategies. Although any one such event occurs infrequently, it may significantly affect outcome in an individual case. Moreover, *different* "rare events" will be encountered often enough to have a substantial effect upon the decisions reached in large numbers of cases. It should be added that even if individual events occur frequently, particular combinations of these events may not be observed often enough to meet the requirements of actuarial prediction.

Furthermore, whatever the theoretical possibilities, there remains a vast multitude of clinical decisions for which no statistical formulas have yet been developed. As Meehl (1957, p. 273) expressed it, "Mostly we will use our heads, because there just aren't any formulas . . ." In this connection, it is important to differentiate between statistical or actuarial procedures, on the one hand, and automated clinical interpretations, on the other. The latter are computerized reports which have been developed for certain tests, such as the DAT (Ch. 13) and the MMPI (Ch.

---

[7] One study, which seemed to favor the clinical method, was later shown not to support this conclusion because of methodological complications (Goldberg, 1968a; see also J. S. Wiggins, 1973, p. 184).

17). They are not actuarial predictions in the sense of being derived statistically from empirical relations. Rather they represent essentially clinicians' interpretations of test data, which are sufficiently standardized to be coded and programmed for computer retrieval. The programs are developed by skilled clinicians, who may draw upon some statistical data from published sources along with personality theory and clinical experience. An account of various ways in which computers may be used in the clinical process is provided by Kleinmuntz (1975).

In summary, the most effective procedure in our present state of knowledge combines clinical and statistical approaches. The clinician should utilize all the objective test data, norms, regression equations, and other actuarial strategies applicable to the particular situation, while supplementing this information with facts and inferences attainable only through clinical methods. The validity of clinical predictions against actual outcomes should be systematically investigated whenever feasible. More data are also needed on the consistency of predictions about the same persons made by different clinicians and by the same clinicians at different times. Insofar as possible, the process and cues on which clinical predictions are based should be made explicit in clinical records. Such a practice would not only facilitate research and training, but would also serve to encourage reliance on sound data and defensible interpretations. Finally, the "clinician as instrument" is an important concept in this connection. Undoubtedly the objectivity and skill with which data are gathered and interpreted—and the resulting accuracy of predictions—vary widely with the abilities, personality, professional training, and experience of individual clinicians.

REPORT WRITING

In Chapter 3, we examined some of the broad issues involved in the communication of test results, with particular reference to ethical and social implications. For the clinical psychologist, such communication usually includes the preparation of a written test report or case report, which may or may not be followed by discussion or consultation with examinee, parents, teachers, or other professionals. Even in those situations that do not require a written report, it is well to prepare such a report as a record for future reference. The preparation of a report also helps to organize and clarify the clinician's own thinking about the case and to sharpen his interpretations. Report writing represents the final stage in the clinician's synthesizing function, described in the preceding section. In its content, the report should draw upon all the data sources (test and nontest) available to the clinician.

Several books and many articles provide guidelines for report writing.

Sattler (1974, Chs. 24, 25) gives a comprehensive view of problems and pitfalls, with many practical suggestions. He also summarizes pertinent research on report writing. Other useful sources include Hammond and Allen (1953), Huber (1961), and Robb, Bernardoni, and Johnson (1972, Ch. 11). These books also contain illustrative case reports. The writing of reports pertaining to specific intelligence tests is discussed and illustrated by Glasser and Zimmerman (1967, Ch. 2) for the WISC; Zimmerman, Woo-Sam, and Glasser (1973, Ch. 15) for the WAIS; and Sattler (1974, Ch. 25) for the Stanford-Binet, WISC, and WPPSI. At the other extreme, an intensive and extensive study of a single case is reported by Holt (1971, Chs. 5–9). This is the case of a man, first studied at the age of 26 with a variety of both standardized tests and informal procedures and then reassessed 26 years later.

Without duplicating the many lists of helpful suggestions that can be found in available sources, we shall focus briefly on some of the major points. First, there is no one standard form or outline for all clinical reports. Both content and style should and do vary with the purpose of the case study, the context in which it is conducted, the recipients to whom the report is addressed, and the theoretical orientation and professional background of the clinician. It is especially important to adapt the report to the needs, interests, and background of those who will receive it. Surveys conducted among the consumers of clinical reports indicate, for example, that teachers welcome concrete recommendations in the reports; psychiatrists, on the other hand, tend to prefer reports concentrating on findings and interpretations, without specific recommendations (see Sattler. 1974, p. 359). Reports intended for both professional and lay readers should contain a brief descriptive summary, followed by a more detailed technical report of specific findings (Hammond & Allen, 1953, Ch. 3).

The content of the report should follow directly from the purpose of the testing or case study, as specified in the reasons for referral or consultation. From the mass of data the clinician has gathered by both formal and informal methods, he should select whatever is relevant to answering the questions raised at the outset. The current trend, moreover, is to make reports action-oriented, i.e., directed toward recommendations or decisions regarding instructional programs, therapy, occupational choice, or other appropriate actions. Diagnosis, etiology, personality description, and understanding of personality dynamics tend to be deemphasized (see, e.g., Arthur, 1969). It should also be recognized, however, that enhancing self-understanding may be an important goal of some case studies and is an integral part of certain types of psychotherapy.

The report should concentrate on each individual's differentiating characteristics—his high and low points—rather than on traits in which his standing is close to the average. A test of the effectiveness of a report is to see whether it is unique to the individual or whether it applies equally well to other persons of the same age, sex, education, socioeconomic level, or other demographic variables. It is a relatively easy task to prepare a pseudo-report from general, stereotyped statements that apply to most people. A considerable body of research has demonstrated that such reports are readily accepted as "remarkably accurate" self-descriptions by a large majority of persons. This illusory validation has been called the "Barnum effect," after Phineas T. Barnum, the famous showman who is credited with the remark that there is a fool born every minute.[8] Reliance on such generally applicable personality descriptions is a favorite device of fortune tellers and other charlatans.

The primary focus of the report should be on interpretations and conclusions, although test records and other detailed data may be separately appended. Specific data, such as individual responses and subtest scores, should ordinarily be cited only to illustrate or clarify a point. The report should be a carefully organized and integrated account. If the data contain inconsistencies of sufficient importance to be included, the clinician should call attention to the inconsistency and if possible suggest reasons for it. In any descriptive statements about the individual's performance, the frame of reference should be made clear. Is the statement based on a criterion-referenced or a norm-referenced evaluation? If the latter, with what norms is the individual being compared? If the comparison is intra-individual, it should be clearly reported as such. For example, one should not write: "His numerical aptitude is superior," but "His numerical aptitude is clearly above the level of his other aptitudes." The clinician's statements should also be so expressed as to reflect his level of certainty for each fact or inference presented.

Many suggestions can be offered regarding writing style.[9] As in all scientific writing, the goal of the clinical report is to communicate, not to obfuscate. Simple, direct writing is not easy to produce. It calls for careful planning, revision, and editing; and it can be improved with practice.

---

[8] The origin of the "Barnum effect" can be traced to a mimeographed report distributed by Donald G. Paterson at the University of Minnesota and subsequently cited by some of his colleagues (Dunnette, 1957; Meehl, 1956). For relevant research, see Forer (1949), Merrens and Richards (1970), Snyder and Larson (1972), Stagner (1958), Sundberg (1955), and Ulrich, Stachnik, and Stainton (1963).

[9] Books on the preparation of clinical reports usually contain helpful hints for good writing, as well as references to standard manuals of style. Special attention is here called to two rather diverse sources: *The Publication Manual of the American Psychological Association* (1974) and *The Elements of Style* by Strunk and White (1972), an entertaining little book which should make writing less painful for both writer and reader.

Avoid the use of technical terms, except when they truly serve to facilitate communication. Do not try to instruct the reader or to impress him with your competence and erudition—that is not the purpose of a clinical report. Write from the reader's point of view. Keep an image of the reader before you as you write. It will keep you from going too far astray.

PART 5

Personality Tests

CHAPTER 17

Self-Report Inventories

ALTHOUGH the term "personality" is sometimes employed in a broader sense, in conventional psychometric terminology "personality tests" are instruments for the measurement of emotional, motivational, interpersonal, and attitudinal characteristics, as distinguished from abilities. In the next four chapters, the major varieties of personality tests will be examined. This chapter will deal with personality inventories. Chapter 18 will consider available techniques for the measurement of interests, attitudes, and values. The instruments to be covered in both of these chapters are essentially paper-and-pencil, self-report questionnaires suitable for group administration. The use of projective techniques for the assessment of personality characteristics will be discussed in Chapter 19. In Chapter 20 we shall survey a number of miscellaneous approaches to the measurement of personality, many of which are still in an experimental stage.

The number of available personality tests runs into several hundred. Especially numerous are the personality inventories and the projective techniques. In this book, we shall be concerned primarily with the types of approaches that have been explored in the assessment of personality. A few of the most widely known tests of each type will be briefly described for illustrative purposes. Many books have been written exclusively about personality assessment through tests as well as through other techniques. For more detailed treatment of the topic, the reader is referred to such specialized books (e.g., Kleinmuntz, 1967; Lanyon & Goodstein, 1971; McReynolds, 1968, 1971, 1975; Megargee, 1966b; Mischel, 1968; I. G. Sarason, 1972; J. S. Wiggins, 1973).

In the development of personality inventories, several approaches have been followed in formulating, assembling, selecting, and grouping items. Among the major procedures in current use are those based on content validation, empirical criterion keying, factor analysis, and personality theory. Each of these approaches will be discussed and illustrated in the following sections. It should be noted, however, that they are not alternative or mutually exclusive techniques. Theoretically, all could be combined in the development of a single personality inventory. In actual practice, several inventories have utilized two or more of these procedures.

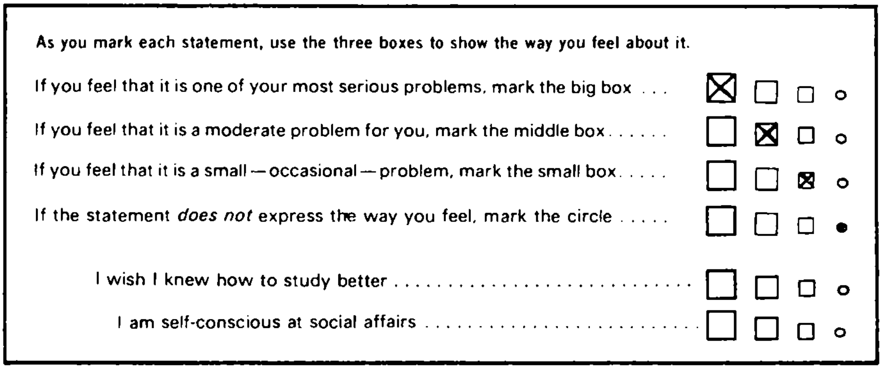
Although some personality tests are used as group screening instruments, the majority find their principal application in clinical and counseling contexts. For this reason, the next four chapters represent largely a continuation of the discussion of clinical testing begun in Chapter 16. In their present state of development, most personality tests should be regarded either as aids in clinical assessment or as research instruments.

CONTENT VALIDATION

The prototype of self-report personality inventories was the Woodworth Personal Data Sheet, developed for use during World War I. This inventory was essentially an attempt to standardize a psychiatric interview and to adapt the procedure for mass testing. Accordingly, Woodworth gathered information regarding common neurotic and preneurotic symptoms from the psychiatric literature as well as through conferences with psychiatrists. It was in reference to these symptoms that the inventory questions were originally formulated. The questions dealt with such behavior deviations as abnormal fears or phobias, obsessions and compulsions, nightmares and other sleep disturbances, excessive fatigue and other psychosomatic symptoms, feelings of unreality, and motor disturbances such as tics and tremors. In the final selection of items, Woodworth applied certain empirical statistical checks, to be discussed in the next section. Nevertheless, it is apparent that the primary emphasis in the construction and use of this inventory was placed on content validity, as indicated in the sources from which items were drawn as well as in the common recognition of certain kinds of behavior as maladaptive.

One of the clearest examples of content validation in a current personality inventory is provided by the Mooney Problem Check List. Designed chiefly to identify problems for group discussion or for individual counseling, this checklist drew its items from written statements of problems submitted by about 4,000 high school students, as well as from case records, counseling interviews, and similar sources. The checklist is available in junior high school, high school, college, and adult forms. The problem areas covered vary somewhat from level to level. In the high school and college forms, they include health and physical development; finances, living conditions, and employment; social and recreational activities; social-psychological relations; personal-psychological relations; courtship, sex, and marriage; home and family; morals and religion; adjustment to school work; the future—vocational and educational; and curriculum and teaching procedure.

Although the number of items checked in each area can be recorded, the Mooney Problem Check List does not yield trait scores or measures of degree of adjustment. Emphasis is on individual items as self-perceived and self-reported problems or sources of difficulty. While no psychometric evaluation of this instrument has been undertaken, evidence has accumulated indicating its effectiveness. Published research shows that, on the average, students check from 20 to 30 problems; these results suggest that the checklist provides good coverage of problems that students are willing to report. Some data on concurrent validity are available from comparisons of contrasted groups whose reported problem frequencies in relevant areas differ in the expected direction.



As you mark each statement, use the three boxes to show the way you feel about it.

If you feel that it is one of your most serious problems, mark the big box . . . ☒ ☐ ☐ ○

If you feel that it is a moderate problem for you, mark the middle box . . . . . . ☐ ☒ ☐ ○

If you feel that it is a small — occasional — problem, mark the small box . . . . . ☐ ☐ ☒ ○

If the statement *does not* express the way you feel, mark the circle . . . . . ☐ ☐ ☐ ●

I wish I knew how to study better . . . . ☐ ☐ ☐ ○

I am self-conscious at social affairs . . . . ☐ ☐ ☐ ○

Fig. 90. Instructions and Two Typical Items from STS Youth Inventory, Form G.

(Reproduced by permission of Scholastic Testing Service.)

Another checklist of needs and problems, suitable for Grades 9 to 12, is the STS Youth Inventory. The 167 items comprising this checklist are grouped under the following rubrics: My School, After High School, About Myself, Getting Along with Others, and Things in General. An ingenious device incorporated into this inventory is the use of response boxes of different sizes to enable the respondent to suggest the magnitude of each problem, as illustrated in Figure 90. In each of the five areas, the manual provides percentile equivalents for the weighted total of problems checked, from which an individual student profile can be plotted. A more useful analysis with this type of checklist, however, focuses on individual items. For this purpose, the manual reports the percentage of respondents in a national sample, by sex and grade, who marked each response alternative for each item. A similar problem checklist, the STS Junior Inventory, is available for Grades 4 to 8.

It should be noted that with all these inventories some efforts have been made toward empirical validation of scores in each problem area.

Few personality tests in use today rest their claims entirely on content validity. The tests cited in this section, however, have relied principally on content validity in the formulation, selection, and grouping of items.

EMPIRICAL CRITERION KEYING

BASIC APPROACH. Empirical criterion keying refers to the development of a scoring key in terms of some external criterion. This procedure involves the selection of items to be retained and the assignment of scoring weights to each response. In the construction of the previously cited Woodworth Personal Data Sheet, some of the statistical checks applied in the final selection of items pointed the way for criterion keying. Thus, no item was retained in this inventory if 25 percent or more of a normal sample answered it in the unfavorable direction. The rationale underlying this procedure was that a behavior characteristic that occurs with such frequency in an essentially normal sample cannot be indicative of abnormality. The method of contrasted groups was likewise employed in the selection of items. Only symptoms reported at least twice as often in a previously diagnosed psychoneurotic group than in a normal group were retained.

Despite some use of such empirical checks, however, content validation rests essentially on a literal or veridical interpretation of questionnaire items. The response to each question is regarded as an index of the actual presence or absence of the specific problem, belief, or behavior described by the question. In empirical criterion keying, on the other hand, the responses are treated as diagnostic or symptomatic of the criterion behavior with which they were found to be associated. In an early description of this approach, Meehl wrote:

> . . . the verbal type of personality inventory is *not* most fruitfully seen as a "self-rating" or self-description whose value requires the assumption of accuracy on the part of the testee in his observations of self. Rather is the response to a test item taken as an intrinsically interesting segment of verbal behavior, knowledge regarding which may be of more value than any knowledge of the "factual" material about which the item superficially purports to inquire. Thus if a hypochondriac says that he has "many headaches" the fact of interest is that he *says this* (Meehl, 1945, p. 9).

A self-report inventory is indubitably a series of standardized verbal stimuli. When criterion-keying procedures have been followed, the responses elicited by these stimuli are scored in terms of their empirically established behavior correlates. They are thus treated like any other psychological test responses. That questionnaire responses may correspond to the subject's *perception* of reality does not alter this situation. It merely provides one hypothesis to account for the empirically established validity of certain items.

THE MINNESOTA MULTIPHASIC PERSONALITY INVENTORY. The outstanding example of criterion keying in personality test construction is the Minnesota Multiphasic Personality Inventory (MMPI). Not only is the MMPI the most widely used personality inventory, but it has also stimulated a flood of research. To date, over 3,500 references have been published about this test. A considerable portion of this research is concerned with factorial analyses of the MMPI scales and with the operation of response styles in its scores, topics to be discussed in later sections of this chapter.

The MMPI consists of 550 affirmative statements, to which the examinee gives the responses: "True," "False," or "Cannot say." In the individual form of the test, the statements are printed on separate cards, which the respondent sorts into the three categories. Later, a group form was prepared, in which the statements are printed in a test booklet and the responses are recorded on an answer sheet. Both forms were designed for adults from about 16 years of age upward, although they have also been employed successfully with somewhat younger adolescents (Hathaway & Monachesi, 1963). Although the card form may be preferable when testing disturbed patients or persons of low educational or intellectual level, the booklet form is now used for most purposes. The MMPI items range widely in content, covering such areas as: health, psychosomatic symptoms, neurological disorders, and motor disturbances; sexual, religious, political, and social attitudes; educational, occupational, family, and marital questions; and many well-known neurotic or psychotic behavior manifestations, such as obsessive and compulsive states, delusions, hallucinations, ideas of reference, phobias, and sadistic and masochistic trends. A few illustrative items are shown below:[1]

I do not tire quickly.

Most people will use somewhat unfair means to gain profit or an advantage rather than to lose it.

I am worried about sex matters.

When I get bored I like to stir up some excitement.

I believe I am being plotted against.

---

[1] Reproduced by permission. Copyright 1943, renewed 1970 by the University of Minnesota. Published by The Psychological Corporation, New York. N.Y. All rights reserved.

In its regular administration, the MMPI provides scores on ten "clinical scales," listed below:

1. Hs: Hypochondriasis
2. D: Depression
3. Hy: Hysteria
4. Pd: Psychopathic deviate
5. Mf: Masculinity-femininity
6. Pa: Paranoia
7. Pt: Psychasthenia
8. Sc: Schizophrenia
9. Ma: Hypomania
0. Si: Social introversion

Eight of these scales consist of items that differentiated between a specified clinical group and a normal control group of approximately 700 persons. The latter were all visitors at the University of Minnesota hospitals, and represented a fairly adequate cross section of the Minnesota population of both sexes between the ages of 16 and 55. The clinical groups varied in size, but most contained about 50 cases. These scales were thus developed empirically by criterion keying of items, the criterion being traditional psychiatric diagnosis. Items for the Masculinity-femininity scale were selected in terms of frequency of responses by men and women. High scores on this scale indicate a predominance of interests typical of the opposite sex. The Social introversion scale, added later, was derived from the responses of two contrasted groups of college students selected on the basis of extreme scores on a test of introversion-extraversion. This scale was also found to be significantly related to the number of extracurricular activities in which high school or college students participated.

A special feature of the MMPI is its utilization of three so-called validity scales.[2] These scales are not concerned with validity in the technical sense. In effect, they represent checks on carelessness, misunderstanding, malingering, and the operation of special response sets and test-taking attitudes. The validating scores include:

Lie Score (L): based on a group of items that make the examinee appear in a favorable light, but are unlikely to be truthfully answered in the favorable direction. (E.g., I do not like everyone I know.)

Validity Score (F): determined from a set of items very infrequently answered in the scored direction by the standardization group. Although representing undesirable behavior, these items do not cohere in any pattern of abnormality. Hence, it is unlikely that any one person actually shows all or most of these symptoms. A high F score may indicate scoring errors, carelessness in responding, gross eccentricity, or deliberate malingering.

Correction Score (K): utilizing still another combination of specially chosen items, this score provides a measure of test-taking attitude, related to both L and F, but believed to be more subtle. A high K score may indicate defensiveness or an attempt to "fake good." A low K score may represent excessive frankness and self-criticism or a deliberate attempt to "fake bad."

---

[2] Originally, there was also a "Question Score," representing the number of items put in the "Cannot Say" category or omitted in the booklet form. Because most examinees omit very few items, this score is no longer regularly used.

The first two scores (L, F) are ordinarily used for an overall evaluation of the test record. If either of these scores exceeds a specified value, the record is considered invalid. The K score, on the other hand, was designed to function as a suppressor variable. It is employed to compute a correction factor which is added to the scores on some of the clinical scales in order to obtain adjusted totals. Because this use of the K score is questionable, however, scores on the affected scales are often reported with and without this correction. An unusally high K score would in itself make the record suspect and call for further scrutiny.

Since the publication of the MMPI in its initial form, about 300 new scales have been developed, most of them by independent investigators who had not participated in the construction of the original test (Dahlstrom, Welsh, & Dahlstrom, 1975). These scales vary widely in the nature and breadth of the criteria against which items were evaluated. Several scales were developed within normal populations to assess personality traits unrelated to pathology. Some scales have subsequently been applied to the test records of the original MMPI normal standardization sample, thus providing normative data comparable to those of the initial clinical scales. Examples of the new scales include Ego Strength (ES), Dependency (Dy), Dominance (Do), Prejudice (Pr), and Social Status (St). Other scales have been developed for highly specific purposes and are more limited in their applicability. Still another grouping of MMPI items is represented by the *content scales* developed by Wiggins (1966b). In the construction of these scales, item clusters based on a subjective classification of content were revised and refined through factor-analytic and internal-consistency procedures. The resulting 13 scales have proved promising in diagnosis and may serve as a useful supplement in the interpretation of the original clinical scales.

In its regular administration, the MMPI now yields 13 scores, including the 9 original clinical scales, the Si scale, and the 3 validating scales. Norms on the original control sample of approximately 700 persons are reported in the form of standard scores with a mean of 50 and an *SD* of 10. These standard scores are used in plotting profiles, as illustrated in Figure 91. Any score of 70 or higher—falling 2 *SD*'s or more above the mean—is generally taken as the cutoff point for the identification of pathological deviations. It should be noted, however, that the clinical significance of the same score may differ from one scale to another. A score of 75 on the Hypochondriasis and on the Schizophrenia scales, for example, may not indicate the same severity of abnormality.

There is considerable evidence to suggest that, in general, the greater

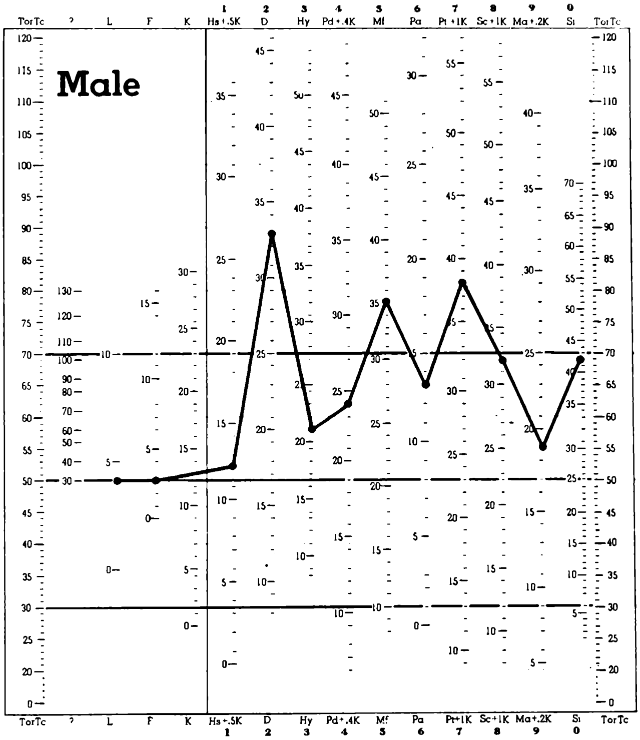


FIG. 91. MMPI Profile of a 34-Year-Old Married Professional Man Seen on an Outpatient Psychiatric Referral.

(Data from Dahlstrom, Welsh, & Dahlstrom, 1972, p. 295. © Copyright 1960, 1972 by the University of Minnesota, published by the University of Minnesota Press, Minneapolis. Profile chart reproduced by permission. Copyright 1948, The Psychological Corporation, New York, N.Y. All rights reserved.)

the number and magnitude of deviant scores on the MMPI, the more likely it is that the individual is severely disturbed. For screening purposes, however, shorter and simpler instruments are available. It is clear that the principal applications of the MMPI are to be found in differential diagnosis. In using the inventory for this purpose, the procedure is much more complex than the labels originally assigned to the scales might suggest. The test manual and related publications now caution against literal interpretation of the clinical scales. For example, we cannot assume that a high score on the Schizophrenia scale indicates the presence of schizophrenia. Other psychotic groups show high elevation on this scale, and schizophrenics often score high on other scales. Moreover, such a score may occur in a normal person. It is partly to prevent possible misinterpretations of scores on single scales that the code numbers 0 to 9 have been substituted for the scale names in later publications on the MMPI.

The original MMPI scales were designed to differentiate between normals and certain traditional diagnostic categories. In subsequent usage, however, the scales have been treated more and more as linear measures of personality traits. These traits have been identified in terms of the salient characteristics of the populations investigated. For example, an elevated score on Scale 6 (originally labeled "Paranoia") may suggest a high degree of resentfulness and suspiciousness. A further complication is that the scales are usually multidimensional. This follows from the fact that the diagnostic groups employed in selecting items generally differed from the normal controls in more than one trait. And some of the same traits may occur in more than one scale, as indicated in part by the presence of common items. It is these two facts—multidimensionality and overlap of MMPI scales—that account for some of the advantages of pattern analysis over the interpretation of single scales. Comparing an individual's scores on two or more scales often helps to identify the trait associated with a particular pattern of scores.

To expedite the interpretation of score patterns, systems of numerical profile coding have been developed. In such codes, the sequence and arrangement of scale numbers show at a glance which are the high and low points in the individual's profile. According to one of the most widely used codes, the profile illustrated in Figure 91 would be coded as follows: 27″5′$\underline{80}$64 – 391/$\underline{\text{KLF}}$/. The scales are listed in descending order of magnitude of their scores; successive 10-point score intervals are separated by "elevation symbols"; scales whose scores are within one point of each other are underlined. The same system is used in recording the three validating scores (Dahlstrom, Welsh, & Dahlstrom, 1972; Hathaway & McKinley, 1967).

As an aid in the diagnostic interpretation of MMPI profiles, Hathaway and Meehl (1951) prepared *An Atlas for the Clinical Use of the MMPI.* This *Atlas* provides coded profiles and short case histories of 968 patients, arranged according to similarity of profile pattern. A similar codebook utilizing data from over 4,000 college students examined in a college counseling center was prepared for use by counselors (Drake & Oetting, 1959). Still another atlas was prepared for use with high school populations (Hathaway & Monachesi, 1961). More recent codebooks have endeavored to simplify and systematize the interpretation of a small number of major profiles, in order to permit actuarial description in "cookbook fashion" (Gilberstadt & Duker, 1965; Marks, Seeman & Haller, 1974). Validation of the MMPI has thus proceeded by the accumulation of empirical data about persons who show each profile pattern or code.

By such a process, the construct validity of each MMPI code is gradually built up. The MMPI *Handbook*, compiled by Dahlstrom, Welsh, and Dahlstrom (1972), contains a comprehensive survey of available interpretive data on major profile patterns.

A further step in the evolution of MMPI score interpretation is the development of several computerized systems for completely automated profile interpretations. Descriptions and critical evaluations of some of these systems can be found in the *Seventh Mental Measurements Yearbook* (Buros, 1972, No. 105–109; see also Butcher, 1969, Chs. 5 & 6; Dahlstrom et al., 1972, Ch. 8). Available services vary widely in several parameters. Some are designed chiefly for screening purposes and are largely descriptive summaries; others provide highly interpretive statements. Some rely on standard published interpretive guidelines; others utilize special keys and extensive research conducted in the development of the system itself. One service provides different levels of reports to suit the needs and qualifications of different recipients. Like all test results, the MMPI computer reports need to be considered in the context of other information about the individual. Recognizing the danger of potential misuses, the American Psychological Association (1966, p. 1141) adopted a set of guidelines for computer-based test-interpretation services. The APA test *Standards* (1974, B6 & J4) also set forth as an essential requirement the reporting of the rationale and evidence on which such interpretive systems are based.

The MMPI record illustrated in Figure 91 was submitted for computer scoring and interpretation to five services and the results are fully reproduced in the MMPI *Handbook* (Dahlstrom et al., 1972, Ch. 8). In Table 31 is the narrative portion of the report prepared by the Mayo Clinic–Psychological Corporation system, the simplest of the available programs. Its reports concentrate on present symptomatology and emotional status. References to diagnosis, prognosis, or etiology, as well as recommendations for specific treatment, are avoided.

Although the misleading psychiatric labels have been dropped in the coded profiles, it should be noted that MMPI items are still grouped into scales on the basis of such obsolescent categories. Factorial analyses based on intercorrelations of items and of scales indicate that items would be differently grouped on the basis of their empirically established interrelations. Furthermore, high intercorrelations among the basic MMPI clinical scales make their value in differential diagnosis questionable. For differential diagnosis, it would have been better to select items by comparing the responses of each clinical group, not with those of normals, but with those of other clinical groups. On the other hand, an extensive body of normative data and clinical experience pertaining to the old scales has accumulated over the years (Dahlstrom et al., 1972; Swenson, Pearson, & Osborne, 1973). In order not to lose this store of information,

TABLE 31
Automated Computer Interpretation of MMPI Profile Reproduced in Figure 91

(From Dahlstrom, Welsh, & Dahlstrom, 1972, p. 309. © Copyright 1960, 1972 by the University of Minnesota, published by the University of Minnesota Press, Minneapolis.)

Severely depressed, worrying, indecisive, and pessimistic

Rigid and meticulous. Worrisome and apprehensive. Dissatisfied with social relationships. Probably very religious and moralistic

Probably sensitive and idealistic with high esthetic, cultural, and artistic interests

Tends toward abstract interests such as science, philosophy, and religion

Probably retiring and shy in social situations

Sensitive. Alive to opinions of others

Independent or mildly nonconformist

Normal energy and activity level

Number of physical symptoms and concern about bodily functions fairly typical for clinic patients

Consider psychiatric evaluation

later efforts have been directed toward the reinterpretation of the old scales in terms of empirically derived profile codes (Butcher, 1972).

A closely related limitation of the MMPI stems from inadequate reliabilities of some of the scales. The effectiveness of any profile analysis is weakened by chance errors in the scores on which it is based. If individual scale scores are unreliable and highly intercorrelated, many of the interscore differences that determine the profile code may have resulted from chance. Retest reliabilities on normal and abnormal adult samples reported in the manual range from the .50's to the low .90's. Other studies, on both normal groups and psychiatric patients, have found even wider variation among scales in retest as well as split-half reliabilities. Some of the split-half reliabilities were especially low, a finding that is not surprising in view of the heterogeneity of item content of some of the scales (Dahlstrom et al., 1972, 1975). It should also be noted that certain scales (notably Scale 2—Depression) assess behavior that is so variable over time as to render retest reliability inappropriate.

Still another limitation of the MMPI pertains to the size and representativeness of the normative sample. The standard scores from which all profile codes are derived are expressed in terms of the performance of the control group of approximately 700 Minneapolis adults tested in the original standardization. Such a normative sample appears quite inadequate when compared, for example, with the nationwide standardization samples employed with many of the ability tests discussed in earlier parts of this book. That the norms may vary appreciably in different normal populations is illustrated by the finding that the means obtained by college students are consistently above 50 on some of the scales (J. H. Clark, 1954; Gilliland & Colgin, 1951; Lanyon, 1968). In one study of 600 college students, 39 percent received scores above 70 in one or more scales (Gilliland and Colgin, 1951). In effect, the original standardization sample should be regarded as a nonnormative fixed reference group, in terms of which the score scale is defined. The much more extensive data subsequently collected with reference to profile codes can then provide all normative interpretation.

Even more than ability tests, personality tests can be expected to show large subcultural as well as cultural differences. As would be anticipated, studies conducted in other countries reveal significant elevation on certain scales when profiles are based on the original Minnesota norms (Rosen & Rizzo, 1961; Sundberg, 1956; Taft, 1957). Similar differences have been found among subcultures in the United States (Baughman & Dahlstrom, 1968; Butcher, Ball, & Ray, 1964; Harrison & Kass, 1967; McDonald & Gynther, 1963; Miller, Knapp, & Daniels, 1968). Any explanation of such cultural and subcultural differences requires specific knowledge of cultural conditions and other circumstances prevailing within each group. As shown in Chapter 12, cultural differentials may operate at many different levels. Group differences in MMPI scores could, for example, reflect nothing more than differences in interpretation of individual items or of instructions. High elevation in some groups could result from strong traditions of self-depreciation and modesty. Cultural differences in the type of behavior considered socially desirable may likewise influence scores. In still other groups, high scores may indicate the prevalence of genuine emotional problems arising from child-rearing practices, conflicts of social roles, minority group frustrations, and other broad cultural differences. In any event, an individual's MMPI profile should be interpreted in the light of information about such demographic variables as age, sex, education, socioeconomic status, ethnic group, and geographic milieu. Some normative data have now been accumulated with reference to these variables (Dahlstrom et al., 1972, 1975; Lanyon, 1968).

In summary, the MMPI is essentially a clinical instrument whose proper interpretation calls for considerable psychological sophistication. If the simplified actuarial interpretations and computer analyses are perceived as diagnostic aids for the overworked clinician, they can serve a useful purpose. There is danger, however, that the trend toward automation may encourage interpretation of MMPI profiles by inadequately trained users.

CALIFORNIA PSYCHOLOGICAL INVENTORY. Besides stimulating a proliferation of scoring scales, the MMPI has served as a basis for the development of other widely used inventories. An outstanding example is the California Psychological Inventory (CPI).[3] While drawing about half of its items from the MMPI, the CPI was developed specifically for use with normal populations from age 13 up. Consisting of 480 items to be answered *True* or *False*, the CPI yields scores in 18 scales. Three are "validity" scales designed to assess test-taking attitudes. These scales are designated as: Sense of well-being (Wb), based on responses by normals asked to "fake bad"; Good impression (Gi), based on responses by normals asked to "fake good"; and Communality (Cm), based on a frequency count of highly popular responses. The remaining 15 scales provide scores in such personality dimensions as Dominance, Sociability, Self-acceptance, Responsibility, Socialization, Self-control, Achievement-via-conformance, Achievement-via-independence, and Femininity.

For 11 of these 15 scales, items were selected on the basis of contrasted group responses, against such criteria as course grades, social class membership, participation in extracurricular activities, and ratings. The ratings were obtained through peer nominations, found to be an effective assessment technique for many interpersonal traits. For the remaining 4 scales, items were originally grouped subjectively and then checked for internal consistency. Cross validation of all scales on sizable samples has yielded significant group differences, although the overlapping of contrasted criterion groups is considerable and criterion correlations are often low.

As in the MMPI, all scores are reported in terms of a standard score scale with a mean of 50 and an *SD* of 10; this scale was derived from a normative sample of 6,000 males and 7,000 females, widely distributed in age, socioeconomic level, and geographic area. In addition, means and *SD*'s of scores on each scale are given for many special groups. Retest as well as internal consistency reliability coefficients of the individual scales compare favorably with those found for other personality inventories (Megargee, 1972). Intercorrelations among scales are relatively high. All but four scales, for example, correlate at least .50 with one or more other scales, indicating considerable redundancy among the 18 scores.

On the whole, however, the CPI is one of the best personality inventories currently available. Its technical development is of a high order and it has been subjected to extensive research and continuous improvement (see Megargee, 1972). In contrast to the clinical interpretation of coded profiles employed with the MMPI, research with the CPI has provided a number of regression equations for the optimal weighting of scales to predict such criteria as delinquency, parole outcome, high school and college grades, and the probability of high school dropout. Cross-cultural studies with individual scales, such as Socialization and Femininity, have also yielded promising validity data against local criteria within different cultures. Extensive research with occupational groups is yielding regression equations and corresponding "derivative profiles" for predicting performance in many fields of work. Computerized profile interpretations are also available. In addition to the basic scales, several additional scales have been developed by various methods.

---

[3] Other examples include the Taylor Manifest Anxiety Scale (J. A. Taylor, 1953) and the Minnesota Counseling Inventory (Berdie & Layton, 1960—listed in Appendix E).

FACTOR ANALYSIS IN TEST DEVELOPMENT

THE GUILFORD INVENTORIES. In the effort to arrive at a systematic classification of personality traits, a number of psychologists have turned to factor analysis. A series of studies by Guilford and his co-workers represents one of the pioneer ventures in this direction (see Guilford, 1959, Ch. 16; Guilford & Zimmerman, 1956). Rather than correlating total scores on existing inventories, these investigators computed the intercorrelations among individual items from many personality inventories. As a by-product of this research, three personality inventories were developed and eventually combined into the Guilford-Zimmerman Temperament Survey. This inventory yields separate scores for the following traits, each score based on 30 different items:

G. *General Activity:* Hurrying, liking for speed, liveliness, vitality, production, efficiency vs. slow and deliberate, easily fatigued, inefficient.

R. *Restraint:* Serious-minded, deliberate, persistent vs. carefree, impulsive, excitement-loving.

A. *Ascendance:* Self-defense, leadership, speaking in public, bluffing vs. submissiveness, hesitation, avoiding conspicuousness.

S. *Sociability:* Having many friends, seeking social contacts and limelight vs. few friends and shyness.

E. *Emotional Stability:* Evenness of moods, optimistic, composure vs. fluctuation of moods, pessimism, daydreaming, excitability, feelings of guilt, worry, loneliness, and ill health.

O. *Objectivity:* Thick-skinned vs. hypersensitive, self-centered, suspicious, having ideas of reference, getting into trouble.

F. *Friendliness:* Toleration of hostile action, acceptance of domination, respect for others vs. belligerence, hostility, resentment, desire to dominate, and contempt for others.

T. Thoughtfulness: Reflective, observing of self and others, mental poise vs. interest in overt activity and mental disconcertedness.

P. Personal Relations: Tolerance of people, faith in social institutions vs. faultfinding, critical of institutions, suspicious, self-pitying.

M. Masculinity: Interest in masculine activities, not easily disgusted, hard-boiled, inhibits emotional expression, little interest in clothes and style vs. interest in feminine activities and vocations, easily disgusted, fearful, romantic, emotionally expressive.

The items in the Guilford-Zimmerman Temperament Survey are expressed in the form of affirmative statements, rather than questions. Most concern the examinee directly. A few represent generalizations about other persons. Three examples are given below:

You start work on a new project with a great deal of enthusiasm	YES	?	NO
You are often in low spirits	YES	?	NO
Most people use politeness to cover up what is really "cut-throat" competition	YES	?	NO

The affirmative item form was chosen in the effort to reduce the resistance that a series of direct questions is likely to arouse. In addition, three verification keys are provided to detect falsification and carelessness of response.

Percentile and standard score norms were derived chiefly from college samples. Attention is called to the desirability of interpreting not only single-trait scores but also total profiles. For example, a high score in Emotional Stability is favorable if coupled with a high General Activity score, but may be unfavorable if it occurs in combination with a low General Activity score. In the latter case, the individual may be sluggish, phlegmatic, or lazy. Split-half reliabilities of separate factor scores range from .75 to .85. Higher reliabilities would of course be desirable for the differential interpretation of individual profiles. Similarly, although an effort was made to obtain independent, uncorrelated trait categories, some of the intercorrelations among the 10 traits are still appreciable. Originally presented only on the basis of its factorial validity, the inventory has subsequently been employed in scattered studies of empirical validity, with varied results.

COMREY PERSONALITY SCALES. A more recent example of a multitrait personality inventory constructed through factor analysis is provided by the Comrey Personality Scales. This inventory yields scores in eight personality scales, designated as follows:

T—Trust vs. Defensiveness

O—Orderliness vs. Lack of Compulsion

C—Social Conformity vs. Rebelliousness

A—Activity vs. Lack of Energy

S—Emotional Stability vs. Neuroticism

E—Extraversion vs. Introversion

M—Masculinity vs. Femininity

P—Empathy vs. Egocentrism

The inventory also includes a Validity Check scale to detect random or inappropriate marking, and a Response Bias scale to assess the tendency to respond in a socially desirable direction.

A special feature of the Comrey inventory is the use of a 7-point scale for recording item responses, which ranges from "always" to "never" or from "definitely" to "definitely not," depending upon the nature of the item. Test construction procedures followed in the selection and grouping of items were of high quality. Split-half reliabilities of scale scores are in the high .80's and .90's. Normative data are limited by the size and restricted nature of the standardization samples; users are encouraged to develop local norms. Validation research against external criteria has yielded some promising results and is still being actively pursued. Studies in several countries suggest considerable cross-cultural stability in the trait categories measured by the Comrey scales.

THE CATTELL INVENTORIES. A somewhat different application of factorial methods to the construction of personality inventories is to be found in the work of Cattell (1946, 1957, 1971, 1973). In the effort to arrive at a comprehensive description of personality, Cattell began by assembling all personality trait names occurring either in the dictionary (as compiled by Allport and Odbert, 1936) or in the psychiatric and psychological literature. This list was first reduced to 171 trait names by combining obvious synonyms. The 171-trait list was then employed in obtaining associates' ratings of a heterogenous group of 100 adults. Intercorrelations and factor analyses of these ratings were followed by further ratings of 208 men on a shortened list. Factorial analyses of the latter ratings led to the identification of what Cattell described as "the primary source traits of personality," a designation that seems to imply more universality and stability of results than appear justified by the antecedent research. It is characteristic of Cattell's approach that he regards factor analysis, not as a data-reduction technique, but as a method for discovering underlying, causal traits.

Factors identified through the correlation of ratings, however, may reflect in part the influence of social stereotypes and other constant errors of judgment, rather than the subjects' trait organization. In fact, other investigators have found the same factors when analyzing ratings given to complete strangers as when analyzing ratings assigned to persons whom the raters knew well (Passini & Norman, 1966). Essentially the same factors were again obtained when college students were asked to rate the similarity in meaning between all possible pairs of words describing bipolar trait scales (D'Andrade, 1965). Apparently factor analysis of ratings may reveal more about the raters than about the ratees.

Cattell maintains that his identification of primary personality traits is corroborated by the findings of other studies by himself and other investigators, using not only ratings but also such techniques as questionnaires and objective tests. Some of the alleged similarities in trait descriptions, however, appear forced and not too convincing (Becker, 1960). It should be recalled that an element of subjectivity is likely to enter into the identification of factors, since the process depends on an examination of those measures or items having the highest loadings on each factor (see Ch. 13). Hence, the cross-identification of factors from separate investigations using different measures is difficult. Despite the extensive research conducted by Cattell and his associates for more than three decades, the proposed traits must be regarded as tentative.

On the basis of their factorial research, Cattell and his co-workers have constructed a number of personality inventories, of which the best known is the Sixteen Personality Factor Questionnaire (16 PF) (Cattell, Eber, & Tatsuoka, 1970). Designed for ages 16 and over, this inventory yields 16 scores in such traits as reserved vs. outgoing, humble vs. assertive, shy vs. venturesome, and trusting vs. suspicious. A "motivational distortion" or verification key is also provided for some of the forms. A computerized, narrative reporting service is available.

Owing to the shortness of the scales, reliabilities of factor scores for any single form of the 16 PF are generally low. Even when two forms are combined, parallel-form reliabilities center in the .50's and retests after a week or less often fall below .80. There is also some question about the factorial homogeneity of items within each scale, as well as the factorial independence of scales (Levonian, 1961). Available information on normative samples and other aspects of test construction is inadequate. Empirical validation data include average profiles for more than 50 occupational groups and about the same number of psychiatric syndromes. "Specification equations" are provided for a number of occupations, in the form of multiple regression equations for predicting an individual's criterion performance from scores on the 16 PF.

Similar inventories have been developed for ages 12 to 18 (Jr.–Sr. High School Personality Questionnaire), 8 to 12 (Children's Personality Questionnaire), and 6 to 8 (Early School Personality Questionnaire). Separate inventories have also been published within more limited areas, including anxiety, introversion-extraversion, and neuroticism. These areas correspond to certain second-order factors identified among correlated first-order factors. The latest addition to the series is the Clinical Analysis Questionnaire, a 28-scale inventory which includes: a shortened version of the 16 PF "in clinical dress," with fewer items per factor, reworded to fit clinical contexts; and 12 pathological scales identified through factor analysis of items from the MMPI and other clinical scales (Cattell et al., 1970, Ch. 14). All of these inventories are experimental instruments requiring further development, standardization, and validation.

Factor analysis provides a technique for grouping personality inventory items into relatively homogeneous and independent clusters. Such a grouping should facilitate the investigation of validity against empirical criteria. It should also permit a more effective combination of scores for the prediction of specific criteria. Homogeneity and factorial purity are desirable goals in test construction. But they are not substitutes for empirical validity.

PERSONALITY THEORY IN TEST DEVELOPMENT

Personality theories have usually originated in clinical settings. The amount of experimental verification to which they have subsequently been subjected varies tremendously from one theoretical system to another. Regardless of the extent of such objective verification, a number of personality tests have been constructed within the framework of one or another personality theory. Clinically formulated hypotheses have been especially prominent in the development of projective techniques, to be considered in Chapter 19. While this approach to test construction has been followed less often for self-report inventories, a few outstanding examples are available.

EDWARDS PERSONAL PREFERENCE SCHEDULE. Among the personality theories that have stimulated test development, one of the most prolific has been the manifest need system proposed by Murray and his associates at the Harvard Psychological Clinic (Murray et al., 1938). One of the first inventories designed to assess the strength of such needs is the Edwards Personal Preference Schedule (EPPS). Beginning with 15 needs drawn from Murray's list, Edwards prepared sets of items whose content appeared to fit each of these needs. The complete list of needs, together with the normal percentile chart employed in plotting EPPS scores, can be found in Figure 92. Examples include the need for Achievement (to do one's best and accomplish something difficult), Deference (to conform to what is expected of one), Exhibition (to be the center of atten-

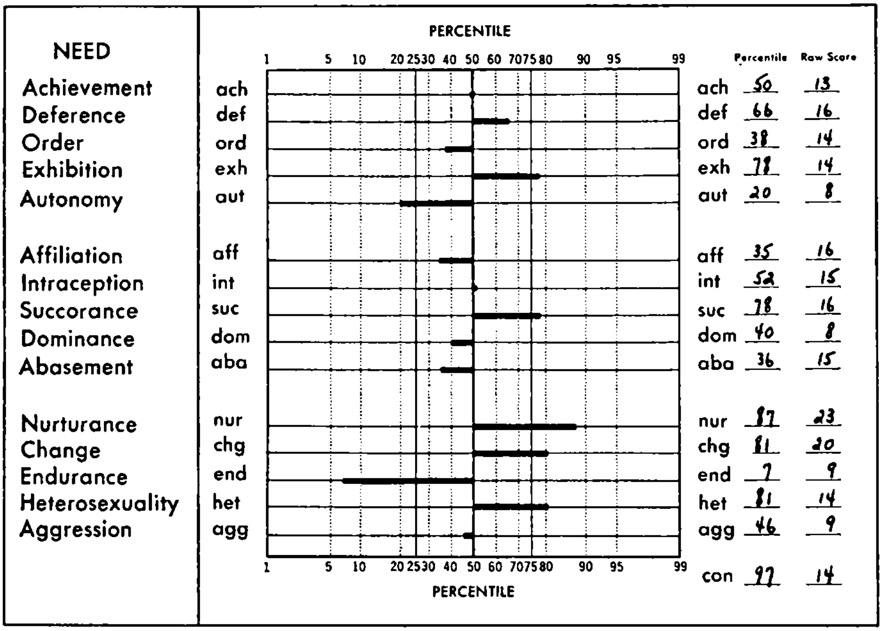


FIG. 92. Profile on the Edwards Personal Preference Schedule, together with List of Needs.

(Reproduced by permission. Copyright 1954, © 1959 by The Psychological Corporation, New York, N.Y. All rights reserved.)

tion), Intraception (to analyze the motives and feelings of oneself and others), Dominance (to influence others and to be regarded as a leader), and Nurturance (to help others in trouble).

The inventory consists of 210 pairs of statements in which items from each of the 15 scales are paired with items from the other 14.[4] Within each pair, the examinee must choose one statement as more characteristic of himself. Two pairs, used as demonstration items, are given below:[5]

A. I like to talk about myself to others.
B. I like to work toward some goal that I have set for myself.

A. I feel depressed when I fail at something.
B. I feel nervous when giving a talk before a group.

The EPPS introduced some ingenious internal checks. To provide an index of respondent consistency, 15 pairs of statements are repeated in identical form. In Figure 92, the last entry (labeled *con* for *consistency*) shows that this respondent made the identical choice in 14 of the 15 pairs, which put him at the 97th percentile of the normative sample in response consistency. Another check yields a profile stability score, which is the correlation between the individual's odd and even scores in the 15 scales.

---

[4] This item form, which represents an important feature of the EPPS, will be discussed further in the next section, as an example of the forced-choice technique.

[5] Reproduced by permission. Copyright 1953, The Psychological Corporation, New York, N.Y. All rights reserved.

The 15 need scores on the EPPS can be evaluated in terms of both percentile and *T*-score norms for college men and college women. These norms were based on 760 men and 749 women tested in 29 colleges scattered over the country. Additional percentile norms are provided from a general adult sample, including 4,031 men and 4,932 women. Drawn from urban and rural areas in 48 states, these respondents constituted a consumer-purchase panel used for market surveys. The need for specific group norms on personality tests is highlighted by the large and significant mean differences found between this consumer panel and the college sample.

It is important to bear in mind that the EPPS employs *ipsative* scores —that is, the strength of each need is expressed, not in absolute terms, but in relation to the strength of the individual's other needs. The frame of reference in ipsative scoring is the individual rather than the normative sample. Because the individual responds by expressing a preference for one item against another, the resulting score is ipsative. Under these conditions, two individuals with identical scores on the EPPS may differ markedly in the absolute strength of their needs. Because of their ipsative nature, the conversion of EPPS scores to normative percentiles may be questioned. The combination of normative and ipsative frames of reference makes interpretation of these scores somewhat confusing and less meaningful than would be the case with a consistently ipsative *or* consistently normative approach.

Retest reliabilities of the 15 scales reported in the manual range from .74 to .88; split-half reliabilities range from .60 to .87. However, both sets of values may be somewhat inflated; the first, through recall of responses over the short interval employed (one week); the second, because of the repetition of identical statements three or four times in different pairs within each scale.

Although the validity data reported in the EPPS manual are meager, a large number of independent validation studies have been published. The results of these studies, however, are often difficult to interpret because most of them failed to take into account the ipsative nature of the scores. With ipsative scores, the mean intercorrelation of individual scales tends to be negative and the mean correlation of all the scales with any outside variable will approach zero (Hicks, 1970). Owing to these artificial constraints, ipsative scores cannot be properly analyzed by the usual correlational procedures. It is not surprising, therefore, that the published validation studies have yielded conflicting and inconclusive results. Although the EPPS introduced several noteworthy features, it is in need of: (a) revision to eliminate certain technical weaknesses, particularly with regard to item form and score interpretation; and (b) properly conducted validation studies utilizing techniques of score pattern analysis appropriate to ipsative scores.

PERSONALITY RESEARCH FORM. The more recently developed Personality Research Form (PRF) reflects many technical advances in test construction, including some item-selection procedures that would have been virtually impossible before computers had reached their present state of development (Jackson, 1967c, 1970). With regard to the procedures followed in defining scales and in formulating and selecting items, the PRF is truly outstanding among personality tests.

The PRF is available in two parallel short forms yielding 14 trait scores and in two parallel longer forms yielding the same 14 scores plus six additional trait scores. There are also two validity scales: an Infrequency Scale in both forms and an additional Desirability Scale in the long form. Designed as an index of carelessness, failure to understand directions, and other nonpurposeful responding, the Infrequency Scale is based on the number of highly unlikely responses chosen by the examinee. Examples of items in this scale include: "I try to get at least some sleep every night" and "I make all my own clothes and shoes." A Desirability Scale is included in the long form, although desirability bias was substantially reduced in advance by the procedures employed in item development and selection. The manual correctly observes that unusually high or low scores on the Desirability Scale may indicate not only atypical test-taking attitudes (e.g., deliberate attempt to create a favorable impression vs. malingering) but also important personality characteristics in their own right (e.g., high self-regard or high degree of conventional socialization vs. low self-regard).

Like several other personality instruments, the PRF took Murray's personality theory as its starting point. Drawing upon the extensive research and theoretical literature that had accumulated during three decades, Jackson formulated behaviorally oriented and mutually exclusive definitions of 20 personality variables. Of these, 12 have the same names as those covered in the EPPS. The additional traits are designated as: Harmavoidance, Impulsivity, Play, Social Recognition, Understanding, Cognitive Structure, Defendence, and Sentience. For each of the 20 personality variables, the manual provides a description of high scorers and a set of defining trait adjectives. Two illustrative scale definitions are shown in Table 32.

Through carefully controlled procedures, a pool of more than 100 items was generated for each scale. Twenty items were then selected for

TABLE 32
Examples of Scale Definitions from Personality Research Form

(From Jackson, 1967c, pp. 6–7; reproduced by permission of Research Psychologists Press, Inc.)

Scale	Description of High Scorer	Defining Trait Adjectives
Cognitive Structure	Does not like ambiguity or uncertainty in information; wants all questions answered completely; desires to make decisions based upon definite knowledge, rather than upon guesses or probabilities.	precise, exacting, definite, seeks certainty, meticulous, perfectionistic, clarifying, explicit, accurate, rigorous, literal, avoids ambiguity, defining, rigid, needs structure.
Sentience	Notices smells, sounds, sights, tastes, and the way things feel; remembers these sensations and believes that they are an important part of life; is sensitive to many forms of experience; may maintain an essentially hedonistic or aesthetic view of life.	aesthetic, enjoys physical sensations, observant, earthy, aware, notices environment, feeling, sensitive, sensuous, open to experience, perceptive, responsive, noticing, discriminating, alive to impressions.

each scale on the basis of high biserial correlation with total scale score and low correlation with scores on other trait scales and on the Desirability Scale. Items yielding extreme endorsement proportions were also eliminated. Through a specially developed computer program, items were assigned to the two parallel forms in terms of biserial correlation with their own scale, as well as endorsement frequency.

Normalized standard scores (T scores) are provided from a sample of about 1,000 male and 1,000 female students from over 30 North American colleges and universities. Although the manual reports that an attempt was made to balance the number of students from different types of colleges and different geographic regions, no data on demographic characteristics of the standardization sample are given. The manual does warn that the norms may be inapplicable to noncollege populations. Various odd-even, Kuder-Richardson, and retest reliability coefficients are reported. In general, the 14 scales in the short forms yielded reliabilities that cluster around .80; reliabilities of the additional six scales in the long form run somewhat lower, dipping as far as the .50's. The manual recommends administration of both parallel forms of each scale if time permits.

The construct validity of the PRF depends to a large extent on the procedures followed in the development and selection of items for each scale. Subsequent factorial analyses corroborated the grouping of items into the 20 scales. Correlations with comparable scales in the California Psychological Inventory and the Guilford-Zimmerman Temperament Survey provided additional support for the identification of the traits. Studies by other investigators, however, indicate that the Murray needs, as measured by such instruments as the PRF, EPPS, TAT (Ch. 19), and ACL (Ch. 20) cannot be assumed to be equivalent (Fiske, 1973; Megargee & Parker, 1968). Data on the empirical validity of the PRS against pooled peer ratings and self-ratings, though still meager, have yielded promising results. The PRS is an excellent research instrument; but more information is needed to determine its effectiveness in practical situations.

In this overview of self-report inventories, the reader has probably noticed an increasing tendency to combine different approaches. This is particularly true of the PRF, which employed all strategies except the empirical criterion keying of items. There is also some evidence suggesting that personality inventory scales constructed by any of the four major methods outlined in this chapter may be equally effective (Hase & Goldberg, 1967). The current trend is to utilize multiple approaches, treating them as different steps in a sequential test-construction strategy (Jackson, 1970, 1973).

TEST-TAKING ATTITUDES AND RESPONSE SETS

FAKING AND SOCIAL DESIRABILITY. Self-report inventories are especially subject to malingering or faking. Despite introductory statements to the contrary, most items on such inventories have one answer that is recognizable as socially more desirable or acceptable than the others. On such tests, the respondent may be motivated to "fake good," or choose answers that create a favorable impression, as when applying for a job or seeking admission to an educational institution. Under other circumstances, he may be motivated to "fake bad," thus making himself appear more psychologically disturbed than he is. This may occur, for example, in the testing of persons on trial for a criminal offense.

Evidence of the success with which examinees can dissemble on personality inventories is plentiful (see, e.g., Jacobs & Barron, 1968; Radcliffe, 1966; Stricker, 1969; J. S. Wiggins, 1966a). A common classroom demonstration consists in asking different groups to assume specified roles. For example, one section of the class is directed to answer each question as it would be answered by a happy and well-adjusted college student; another section is told to respond in the manner of a severely maladjusted person; and in the last section respondents are instructed to answer the items truthfully with reference to their own behavior. Or the same persons may take the test twice, first with instructions to simulate in a specified way and later under ordinary conditions. The results of such studies clearly demonstrate the facility with which the desired impression can be deliberately created on such inventories.

It is interesting to note that specific simulation for a particular vocational objective can also be successfully carried out. Thus, in one study (Wesman, 1952), the responses of the same group of students were compared on two administrations of a personality inventory taken a week apart. On the first testing, the subjects were instructed to pretend they were applying for the position of salesman in a large industrial organization and to answer in a manner designed to increase their chances of employment. On the second testing, the same instructions were given, but the position of librarian was substituted for that of salesman. When the responses were scored for the trait of self-confidence, a conspicuous difference was found in the distributions of scores on the two occasions, the simulated-salesman scores being much higher than the corresponding librarian scores. That job applicants do in fact fake personality test responses was demonstrated in another study (R. F. Green, 1951), in which the scores obtained by a group of applicants were compared with the scores of a comparable group of job holders who were tested for research purposes only. Under these contrasting motivating conditions, the scores of the two groups differed in the expected direction.

The tendency to choose socially desirable responses on a self-report inventory need not indicate deliberate deception on the part of the respondent. A. L. Edwards (1957a), who first investigated the social desirability (SD) variable, conceptualized it primarily as a façade effect or tendency to "put up a good front," of which the respondent is largely unaware. This tendency may indicate lack of insight into one's own characteristics, self-deception, or an unwillingness to face up to one's limitations. Other investigators (Crowne & Marlow, 1964; N. Frederiksen, 1965) have presented evidence to suggest that the strength of the social desirability response set is related to the individual's more general need for self-protection, avoidance of criticism, social conformity, and social approval. On the other hand, the individual who chooses unfavorable items in a self-description may be motivated by a need for attention, sympathy, or help in meeting his personal problems. A person seeking psychotherapy, for example, is likely to make himself appear more maladjusted on a personality inventory than he actually is.

It cannot be assumed that basic research is free from the effects of response sets. In a provocative survey of studies on attitude change, for example, Silverman and Shulman (1970) described a number of such response sets. The results of these investigations may be influenced by such conditions as the subject's perception of what the experimenter expects, his desire to protect his own image, and his wish to please or to frustrate the experimenter. Unsuspected differences in these response sets may account in part for failures to replicate results when the experiments are repeated.

In order to investigate the contribution of the social desirability variable to personality test responses, Edwards (1957a) developed a special social desirability scale. Beginning with 150 heterogeneous MMPI items taken from the three validating keys and the Taylor Manifest Anxiety Scale, Edwards selected 79 items that yielded complete agreement among 10 judges with regard to the socially desirable response. Through item analyses against total scores on this preliminary scale, he shortened the list to 39 items. The SD scale correlated .81 with the K scale of the MMPI, partly because of common items between the two scales. Individual scores on this scale can be correlated with scores on any personality test as a check on the degree to which the social desirability variable has been ruled out of test responses. Whatever the cause of the relation, Edwards argued that insofar as social desirability is correlated with test scores, the effectiveness of the test in discriminating individual differences in specific, content-related traits is reduced.

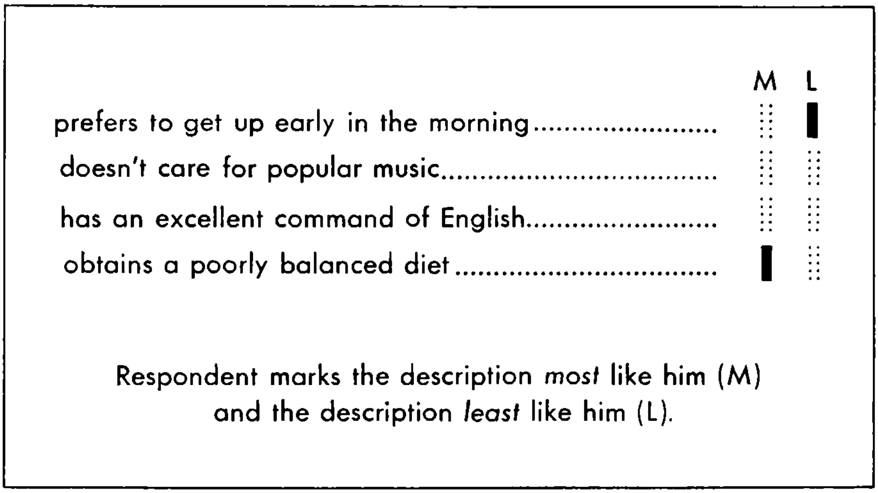
Several procedures have been followed in the effort to meet the problem of faking and related response sets in personality inventories. The construction of relatively "subtle" or socially neutral items may reduce the operation of these factors in some inventories.[6] In a number of situations, the test instructions and the establishment of rapport may motivate the individual to respond frankly, if he can be convinced that it is to his own advantage to do so. This approach would be ineffective in certain situations, however; and it would not have much effect on social desirability response sets of which the individual is unaware. Other attempted solutions include verification keys that detect faking or response sets, such as the F scale of the MMPI, and correction terms, such as those provided by the K scale of the MMPI. Still another procedure, directed not to the detection but to the prevention of dissimulation, is the use of forced-choice items.

FORCED-CHOICE TECHNIQUE. The forced-choice technique was simultaneously developed by several psychologists working in industry or in the armed services during the decade of the 1940s (Jurgensen, 1944; Shipley, Gray, & Newbert, 1946; Sisson, 1948). Essentially, it requires the respondent to choose between two descriptive terms or phrases that appear equally acceptable but differ in validity. The paired phrases may both be desirable or both undesirable. The two demonstration items from the EPPS, reproduced earlier in this chapter, illustrate this item form. Forced-choice items may also contain three, four, or five terms, as illustrated by the sample item from the Gordon Personal Inventory reproduced in Figure 93. In such cases, the respondent must indicate which phrase is most characteristic and which least characteristic of himself. Still another variant requires a choice between two contrasting responses within the same trait, scored for a single scale. Although rarely employed in personality inventories, this item form has the advantage of yielding normative rather than ipsative scores and hence imposing no artificial constraints on the interrelationships among different scales. An example is provided by the Myers-Briggs Type Indicator (Myers, 1962).

---

[6] See, e.g., the development by Block (1965, Ch. 7) of an Ego-Resiliency (Subtle) scale for the MMPI. Jackson (1971), however, has argued that so-called subtle items may be merely items of low validity for the dimension under consideration.

Use of the forced-choice technique to control social desirability requires two types of information regarding each response alternative, viz., its social desirability or "preference index" and its validity or "discriminative index." The latter may be determined on the basis of any specific criterion the inventory is designed to predict, such as academic achievement or success on a particular kind of job; or it may be based on the factor loading of items or their theoretical relevance to different traits. Social desirability can be found by having the items rated for this variable by a representative group, or by ascertaining the frequency with



	M	L
prefers to get up early in the morning........................	⁞	▮
doesn't care for popular music..................................	⁞	⁞
has an excellent command of English........................	⁞	⁞
obtains a poorly balanced diet.................................	▮	⁞

Respondent marks the description *most* like him (M)
and the description *least* like him (L).

Fig. 93. Sample Forced-Choice Item from Gordon Personal Inventory.
(Reproduced by permission of Harcourt Brace Jovanovich, Inc.)

which the item is endorsed in self-descriptions. In a series of studies on many different groups, Edwards (1957a) showed that frequency of choice and judged social desirability correlate between .80 and .90. In other words, the *average* self-description of a population agrees closely with its average description of a desirable personality. Moreover, the rated SD of items remains remarkably stable in groups differing in sex, age, education, socioeconomic level, or nationality. Consistent results were also obtained when the judgments of hospitalized psychiatric patients were compared with those of normal groups.

In order to rule out the operation of the SD variable in the EPPS, Edwards relied exclusively on the forced-choice type of item. Independent research with this test since its publication, however, indicates that the influence of SD may have been reduced but was certainly not eliminated. When EPPS items were presented in a free-choice format, the scores correlated quite highly with the scores obtained with the regular forced-choice form of the test (Lanyon, 1966). The median correlation was .73, several of the correlations approaching the test-retest reliability of the scales. There is evidence that the context in which an item is encountered affects its perceived social desirability. Thus, the SD of an item rated separately may change when that item is paired with another item in a forced-choice format. Not only were significant differences in SD found between the paired items in the EPPS, but the redetermined SD values also yielded substantial correlations with the frequency of item endorsement (Corah et al., 1958). These correlations sometimes approached the correlations found between SD ratings and frequency of endorsement of items presented singly and marked "True" or "False" (A. L. Edwards 1957a, 1966). There is also evidence that forced-choice tests are no more valid than single-stimulus tests (Scott, 1968).

It has been shown, furthermore, that EPPS responses *can* be deliberately faked to create the desired impressions, especially for specific purposes (Borislow, 1958; Dicken, 1959). It cannot be assumed that judged SD is constant for all purposes, even though the judgments of general social desirability obtained from different populations may agree. The relative desirability of the same items for salesmen or for physicians, for example, may differ from their desirability when judged in terms of general cultural norms. Thus, a forced-choice test whose items were equated in general social desirability could still be faked when taken by job applicants, candidates for admission to professional schools, and other specifically oriented groups. From another angle, it has been found that when items are paired on the basis of average *group* judgments of general social desirability, they may be far from equated for *individuals* (N. Wiggins, 1966).

In conclusion, it appears that the forced-choice technique has not proved as effective as had been anticipated in controlling faking or social desirability response sets. At the same time, the forced-choice item format, particularly when it yields ipsative scores, introduces other technical difficulties and eliminates information about absolute strength of individual characteristics that may be of prime importance in some testing situations.

RESPONSE SETS AND RESPONSE STYLES. The tendency to choose socially desirable response alternatives is only one of several response sets that have been identified in test responses. Although the voluminous literature on the operation of response sets in personality inventories dates largely from midcentury, the influence of response sets in both ability and personality tests was observed by earlier investigators (see Block, 1965, Ch. 2). One of the response sets that attracted early attention was *acquiescence,* or the tendency to answer "True" or "Yes." Acquiescence is conceptualized as a continuous variable; at one end of the scale are the consistent "Yeasayers" and at the other end the consistent "Naysayers" (Couch & Keniston, 1960). The implications of this response set for the construction of personality inventories is that the number of items in which a "Yes" or "True" response is keyed positively in any trait scale should equal the number of items in which a "No" or "False" response is keyed positively. This balance can be achieved by the proper selection or rewording of items, as was done in the PRF.

A third response set that has been widely investigated is *deviation,* or the tendency to give unusual or uncommon responses. Berg (1967), who proposed this deviation hypothesis, has argued strongly for the content-free nature of this response set. Accordingly he demonstrated its operation with nonverbal content in a specially developed test requiring an expression of preference for geometric figures.

Research on response sets such as social desirability, acquiescence, and deviation has passed through two principal stages. When first identified, response sets were regarded as a source of irrelevant or error variance to be eliminated from test scores. Efforts were therefore made to rule out their influence through the reformulation of items, the development of special keys, or the application of correction terms. Later these response sets came to be considered as indicators of broad and durable personality characteristics that were worth measuring in their own right (Jackson & Messick, 1958, 1962; J. S. Wiggins, 1962). At this stage, they were commonly described as *response styles.*[7] We have already noted the association of the social desirability response set with certain general behavior tendencies to seek social approval. Evidence has likewise been gathered

[7] This distinction between response sets and response styles is not universal, however. Some writers use the two terms in a different sense (see, e.g., Rorer, 1965).

to suggest pervasive personality differences between Yeasayers and Naysayers (Couch & Keniston, 1960) and between those who choose common responses and those who choose deviant responses (Berg, 1967).

Now we have come full circle. Personality inventory responses are again regarded to have broad diagnostic significance, but in terms of their stylistic properties rather than in terms of specific item content. At the same time the elaborate edifice that has been built around response styles shows signs of crumbling. The empirical data adduced in support of a response-style interpretation of scores on such inventories as the MMPI has been challenged from many directions (Block, 1965; Heilbrun, 1964; Rorer, 1965). Factorial analyses of the MMPI scales by several investigators have generally yielded two major factors that account for nearly all the common variance among the scales. The exponents of response sets and response styles have interpreted these two factors as social desirability and acquiescence, although differing in the relative importance they attribute to each. In a sophisticated statistical analysis of MMPI scales and items, Block (1965) presented strong evidence supporting a content-oriented interpretation of these two factors and indicating that the contribution of response styles to the variance of MMPI scores is negligible. He also demonstrated that the evidence advanced earlier in support of stylistic interpretations of these factors reflected methodological artifacts.

The controversy over response sets and content-versus-style in personality assessment is far from settled (Bentler, Jackson, & Messick, 1971, 1972; Block, 1967, 1971b, 1972; Jackson, 1967a, 1967b; Samelson, 1972). Nevertheless, it may prove to be a tempest in a teapot. Like many scientific controversies, it has stimulated extensive research and has produced several hundred publications. And like many scientific controversies, its net effect will probably be to sharpen our understanding of methodological problems and thereby improve both the construction of personality inventories and the research conducted with them in the future. It is likely that some stylistic scales may ultimately prove to be valid predictors of important personality traits; but it is unlikely that stylistic scales will generally replace content-related scales in personality inventories.

SITUATIONAL SPECIFICITY

Since the 1960s there has been a growing emphasis on situational specificity in personality assessment, together with a tendency to reject personality description in terms of broad traits (Bandura, 1969; Bandura & Walters, 1963; Goldfried & Kent, 1972; Mischel, 1968, 1969, 1973). Criticism has been directed chiefly against the view of personality traits as fixed, unchanging, underlying causal entities. This type of criticism had been anticipated in the earlier research and writings of several psychologists, with regard to all traits—cognitive as well as noncognitive (see Ch. 13).

Situational specificity in particular, however, is much more characteristic of personality traits than it is of abilities. For example, a person might be quite sociable and outgoing at the office, but shy and reserved at social gatherings. Or a student who cheats on examinations might be scrupulously honest in money matters. An extensive body of empirical evidence has been assembled by Mischel (1968) and D. Peterson (1968) showing that individuals do exhibit considerable situational specificity in many nonintellective dimensions, such as aggression, social conformity, dependency, rigidity, honesty, and attitudes toward authority. La Fave's concept of *attitude shifting* is essentially an application of situational specificity to the assessment of attitudes, although developed within a different theoretical orientation from that of Mischel and other social-learning theorists (La Fave, 1974; La Fave, Haddad, & Marshall, 1974).

Part of the explanation for the higher cross-situational consistency and temporal stability of cognitive as compared to noncognitive functions may be found in the greater standardization of the individual's reactional biography in the intellectual than in the personality domain (Anastasi, 1948; 1958a, Ch. 11; 1970). The formal school curriculum, for example, contributes to the development of broadly applicable cognitive skills in the verbal and numerical areas. Personality development, on the other hand, occurs under far less uniform conditions. In the personality domain, moreover, the same response may lead to social consequences that are positively reinforcing in one situation and negatively reinforcing in another. The individual may thus learn to respond in quite different ways in different contexts (Mischel, 1968). Such dissimilarities in experiential history, across individuals as well as across situations, also lead to greater ambiguity in personality test items than is found in cognitive test items. Thus, the same response to a given question on a personality inventory may have a different diagnostic significance from one person to another.

With regard to assessment techniques, several developments since 1960 have encouraged the substitution of specifically identified "behavior of interest" for more broadly defined traits. In the cognitive domain, this focus is illustrated by individualized instructional programs and criterion-referenced testing (Chs. 4 and 14) and by the diagnosis and treatment of learning disabilities (Ch. 16). In the noncognitive or personality domain, the strongest impetus toward a high degree of specificity in testing has come from the procedures variously designated as "social learning," "behavior therapy," and "behavior modification." Essentially, this approach consists in the systematic application of the principles of both classical and operant conditioning to education, remediation, and therapy (Bandura, 1969; Bandura & Walters, 1963; Franks, 1969; Krasner, 1971; Mischel, 1968; Rotter, 1964; Wolpe, 1973). Its roots can be found in early experiments by John B. Watson and his co-workers (see M. C. Jones, 1974; Watson & Watson, 1921) and its subsequent development can be traced through several decades of basic research on learning and personality. Its principal applications occur in clinical psychology and in education.

Specific examples of testing techniques developed for use in behavior-modification programs will be cited in Chapter 20. In the present connection, we shall consider two illustrations of assessment models based on social-learning theory, with its emphasis on the interaction of persons and situations. Endler, Hunt, and Rosenstein (1962) prepared the S-R Inventory of Anxiousness, which identifies different stimulus situations and different modes of response. The authors point out that individuals may differ in the number and variety of situations arousing anxiety, as well as in the number, kind, intensity, and duration of their responses to anxiety-provoking stimuli. In an effort to assess at least some of these parameters, they constructed a questionnaire describing 11 situations, for each of which the respondent rates the intensity of each of 14 reactions (e.g., enjoy the challenge, experience nausea). Later, five additional forms of the questionnaire were prepared.

Analyses of data obtained with the S-R Inventories of Anxiousness (Endler & Hunt, 1966, 1969) indicated that the overall contribution of situations and of individual differences were about equally small. The major part of the variance was attributable to interactions between situations, responses, and individuals. For example, the relative degree of anxiety evoked by different situations varied markedly among persons; and the type of response differed from one person to another and from one situation to another. The original data were later submitted to a multimode factor analysis (see Tucker, 1966), which permitted the identification of three situational factors (interpersonal stress, inanimate danger, novel or unknown situation); three response modes (distress, exhiliration, autonomic responses); and three person or type factors corresponding to three patterns of relationship between situations and response modes. A similar inventory was subsequently developed for the assessment of hostility (Endler & Hunt, 1968).

Another assessment model that incorporates situational specificity is exemplified by the Survey of Problematic Situations, developed by Goldfried and D'Zurilla (1969). This inventory was designed to assess competence among college freshmen, defined as "the effectiveness or adequacy with which an individual is capable of responding to the various problematic situations which confront him" (p. 161). Following an approach designated as behavior-analytic, the authors first assembled a list of over 600 problem situations, derived principally from student reports and supplemented with observations by resident assistants, faculty, and staff members, and with examination of clinical folders from the student counseling center. Through a stepwise process, the list was eventually reduced to 87 problem situations by eliminating redundant and trivial entries, infrequently reported situations, and those showing inadequate response variability. The situations were about equally divided among three categories: academic, interpersonal, and combinations of the two.

In order to develop a scoring scale, the final form of the inventory was administered to a representative sample of freshmen, with the instructions that they try to imagine themselves in each situation and describe in detail their most likely reaction. These responses were then rated by specially chosen judges (e.g., faculty, dormitory counselors) for degree of effectiveness in resolving the problem. In the application of the inventory, each response is scored, as in a product scale, by assigning to it the rating of the sample response that matches it most closely. This will be recognized as essentially a criterion-referenced evaluation. The same general testing model has been followed by other investigators in developing instruments for assessing interpersonal skills in a training program with psychiatric inpatients (Goldsmith, 1973) and for measuring assertion (MacDonald, 1974).[8]

EVALUATION OF PERSONALITY INVENTORIES

It should now be apparent that the construction and use of personality inventories are beset with special difficulties over and above the common problems encountered in all psychological testing. The question of faking and malingering is far more acute in personality measurement than in aptitude testing. The behavior measured by personality tests is also more changeable over time than that measured by tests of ability. The latter fact complicates the determination of test reliability, since random temporal fluctuations in test performance are likely to become confused with broad, systematic behavioral changes. Even over relatively short intervals, it cannot be assumed that variations in test response are restricted to the test itself and do not characterize the area of nontest behavior under consideration. A related problem is the greater situational specificity of responses in the noncognitive than in the cognitive domain.

To a large extent, such problems are shared by all types of personality tests. But for the present we shall limit our discussion to self-report inventories. The acknowledged deficiencies of current personality inventories may be met in at least two major ways. First, personality inventories may be recognized as intrinsically crude instruments and their application restricted accordingly. Second, various procedures for improving the inventories may be explored. Most psychologists today would probably accept some combination of the two approaches, although a few may align themselves exclusively behind one or the other.

---

[8] Personal communication from M. R. Goldfried, September 9, 1974.

A specific illustration of the first approach is provided by the use of a personality inventory merely as a springboard for a clinical interview. In such cases, the interviewer might not even score the inventory in the standard manner, but might simply examine the subject's answers with a view to identifying problem areas for further probing during the interview. Other current practices stemming from a recognition of the pitfalls presented by personality inventories pertain to the interpretation of "poor" versus "good" scores, and the use of inventories in counseling versus selection. In most situations a "poor," or unfavorably deviant, score is likely to signify maladjustment, while a "good" score may be ambiguous. It is also evident that the motivation to create a favorable impression is much stronger in the job applicant than in the person seeking help from a counselor or in the volunteer for a research project. Even in the latter situations, however, complete candor cannot be assumed, because of the prevalence of rationalizations, defensive reactions, and other façade effects. Self-disclosure, together with its correlates, has been a subject of considerable psychological research in its own right (Cozby, 1973).

Attempts to improve self-report inventories by direct attack on the major sources of difficulty have been described throughout this chapter. Among such efforts may be mentioned the application of factor analysis as a means of arriving at more systematic trait categories, the keying of individual items against highly specific criteria, the use of a forced-choice technique, and the development of verification and correction scales. With regard to temporal instability, a promising approach is illustrated by the State-Trait Anxiety Inventory (STAI). This questionnaire was developed in two forms: in one, the examinee is instructed to respond as he *generally* feels ("Trait anxiety"); in the other, as he feels *at the moment* ("State anxiety"). It is noteworthy that retest reliabilities of the trait scale were satisfactorily high (.73 to .86); but for the state scale, for which temporal stability is not expected, they were quite low (.16 to .54). Kuder-Richardson reliabilities, on the other hand, were high for both scales (.83 to .92).

Procedures for dealing with the problem of situational differences and person × situation interactions were illustrated in the preceding section. Social-learning theory and behavior-analytic approaches are undoubtedly exerting a salutary influence on personality assessment as a whole. Their focus on direct sampling of the behavior of interest, their exposure of the circular reasoning involved in hypostatizing traits as stable, underlying causal entities, and their rejection of remote and untestable psychoanalytic (or "psychodynamic") explanations of behavior are helping to clear away some psychometric cobwebs.

A possible danger, however, is that, in the zeal to eradicate fallacious thinking, sound and useful concepts may also be lost. The occasional proposal that diagnostic testing and trait concepts be completely discarded, for example, indicates an unnecessarily narrow definition of both terms. Diagnosis need not imply the labeling of persons, the use of traditional psychiatric categories, or the application of the medical, "disease" model. Diagnostic testing should be used as an aid in describing and understanding the individual, identifying his problems, and reaching appropriate action decisions. Similarly, traits refer to the categories into which behavior must necessarily be classified if we are to deal with it at all—in science or in any other context. The optimum category breadth will vary with the specific purpose of assessment. The hierarchical trait models presented in Chapter 13 can be employed equally well in the description of noncognitive behavior. Under certain circumstances, relatively broad traits will serve best; under other circumstances, very narrow, specifically defined behaviors will need to be assessed.

CHAPTER 18

Measures of Interests, Attitudes, and Values

THE STRENGTH and direction of the individual's interests, attitudes, motives, and values represent an important aspect of his personality. These characteristics materially affect his educational and vocational adjustment, his interpersonal relations, the enjoyment he derives from his avocational pursuits, and other major phases of his daily living. Although certain tests are specifically directed toward the measurement of one or another of these variables, the available instruments cannot be rigidly classified according to such discrete categories as interests, attitudes, and values. Overlapping is the rule. Thus, a questionnaire designed to assess the relative strength of different values, such as the practical, aesthetic, or intellectual, may have much in common with interest inventories. Similarly, such a questionnaire might be said to gauge the individual's attitudes toward pure science, art for art's sake, practical applications, and the like.

The study of *interests* has probably received its strongest impetus from vocational and educational counseling. To a slightly lesser extent, the development of tests in this area has also been stimulated by vocational selection and classification. From the viewpoint of both the worker and the employer, a consideration of the individual's interests is of practical significance. Achievement is a resultant of aptitude and interest. Although these two variables are positively correlated, a high level in one does not necessarily imply a superior status in the other. An individual may have sufficient aptitude for success in a certain type of activity—educational, vocational, or recreational—without the corresponding interest. Or he may be interested in work for which he lacks the prerequisite aptitudes. A measure of both types of variables thus permits a more effective prediction of performance than would be possible from either alone.

The assessment of *opinions and attitudes* originated largely as a problem in social psychology. Attitudes toward different groups have obvious implications for intergroup relations. Similarly, the gauging and prediction of public opinion regarding a wide variety of issues, institutions, or practices are of deep concern to the social psychologist, as well as to the practical worker in business, politics, and other applied fields. The measurement of opinions and attitudes has also made rapid strides in the areas of market research and employee relations.

In this chapter, we shall examine typical standardized tests designed to measure interests, attitudes, values, and related aspects of personality. As in the preceding chapter, attention will be focused on the paper-and-pencil self-report inventory. The majority of interest and attitude measures in current use are of this type. It should be noted, however, that in this area—as in the measurement of all personality characteristics—other approaches are being increasingly explored. A consideration of noninventory techniques will be reserved for Chapters 19 and 20.

INTEREST INVENTORIES

It would seem that the most expedient and direct way of determining an individual's interests in different types of work, educational curricula, or recreational activities would be simply to ask him. But early investigators soon discovered that answers to direct questions about interests are often unreliable, superficial, and unrealistic (see Fryer, 1931, Ch. 5). This is particularly true of children and young people at the ages when information regarding interests is especially useful for counseling purposes.

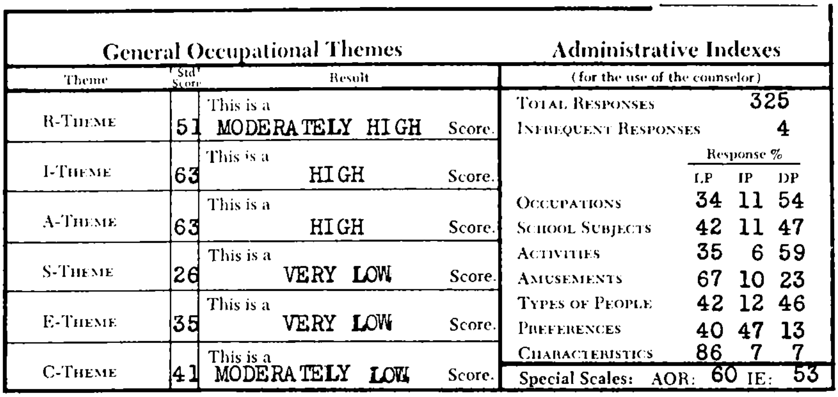
The reasons for this situation are not hard to find. In the first place, most persons have insufficient information about different jobs, courses of study, and other activities. They are thus unable to judge whether they would really like all that their choice actually involves. Their interest—or lack of interest—in a job may stem from a limited notion of what the day-by-day work in that field entails. A second, related factor is the prevalence of stereotypes regarding certain vocations. The life of the average doctor, lawyer, or engineer is quite unlike the versions encountered in movies, television, and popular magazines. The problem, therefore, is that individuals are rarely in a position to know their own interests in various fields prior to actual participation in those fields. And by the time they have had the benefit of such personal contact, it may be too late to profit from the experience, since a change may be too wasteful.

For this reason, it was soon realized that more indirect approaches to the determination of interests would have to be explored. One of the most fruitful of these approaches originated in a graduate seminar on interests conducted at the Carnegie Institute of Technology during the academic year 1919–1920 (D. P. Campbell, 1971, Ch. 11; Fryer, 1931, Ch. 3). Several standardized interest inventories were subsequently prepared as a result of the work begun by their authors while attending this seminar. But the one whose development has been carried furthest is the Strong Vocational Interest Blank (SVIB), constructed by E. K. Strong, Jr. Unlike other early tests, the SVIB has undergone continuing research, revision, and extension.

The interest inventories developed by the Carnegie group introduced two principal procedural innovations. First, the items dealt with the respondent's liking or dislike for a wide variety of specific activities, objects, or types of persons that he has commonly encountered in daily living. Second, the responses were empirically keyed for different occupations. These interests inventories were thus among the first tests to employ criterion keying of items, subsequently followed in the development of such personality inventories as the MMPI and CPI (Ch. 17). It was found that persons engaged in different occupations were characterized by common interests that differentiated them from persons in other occupations. These differences in interests extended not only to matters pertaining directly to job activities, but also to school subjects, hobbies, sports, types of plays or books the individual enjoyed, social relations, and many other facets of everyday life. It thus proved feasible to question the individual about his interests in relatively familiar things, and thereby determine how closely his interests resembled those of persons successfully engaged in different vocations.

STRONG VOCATIONAL INTEREST BLANK. The 1974 form of the Strong Vocational Interest Blank (SVIB), designated as the Strong-Campbell Interest Inventory (SCII), introduced more extensive innovations than did any earlier revision. The principal changes include the merging of the earlier men's and women's forms into a single test booklet and the introduction of a theoretical framework to guide the organization and interpretation of scores. The inventory consists of 325 items grouped into seven parts. In the first five parts, the examinee records his preferences by marking L, I, or D to indicate "Like," "Indifferent," or "Dislike." Items in these five parts fall into the following categories: occupations, school subjects, activities (e.g., making a speech, repairing a clock, raising money for charity), amusements, and day-to-day contact with various types of people (e.g., very old people, military officers, people who live dangerously). The remaining two parts require the respondent to express a preference between paired items (e.g, dealing with things vs. dealing with people) and marking a set of self-descriptive statements "Yes," "No," or "?."

The SCII can be scored only by computer, through several designated scoring agencies. Figures 94, 95, and 96 show sections of the two-page profile booklet on which the different types of scores are reported. For illustrative purposes, scores have been inserted from the profile of a 27-



General Occupational Themes

Theme	Std Score	Result	
R-THEME	51	This is a MODERATELY HIGH	Score.
I-THEME	63	This is a HIGH	Score.
A-THEME	63	This is a HIGH	Score.
S-THEME	26	This is a VERY LOW	Score.
E-THEME	35	This is a VERY LOW	Score.
C-THEME	41	This is a MODERATELY LOW	Score.

Administrative Indexes

(for the use of the counselor)

TOTAL RESPONSES 325

INFREQUENT RESPONSES 4

	Response %		
	LP	IP	DP
OCCUPATIONS	34	11	54
SCHOOL SUBJECTS	42	11	47
ACTIVITIES	35	6	59
AMUSEMENTS	67	10	23
TYPES OF PEOPLE	42	12	46
PREFERENCES	40	47	13
CHARACTERISTICS	86	7	7

Special Scales: AOR: 60 IE: 53

FIG. 94. Portions of SCII Profile, Showing Scores on Six General Occupational Theme Scales (Realistic, Investigative, Artistic, Social, Enterprising, Coventional), Administrative Indexes (used to detect carelessness and test-taking response sets), and Special Scales (Academic Orientation and Introversion-Extroversion).

(Form reprinted with permission of the publisher from *Manual for the SCII*, Form T325 of the STRONG VOCATIONAL INTEREST BLANK, by David P. Campbell [Stanford University Press, 1974], p. 18. Scores in Fig. 94, 95, and 96 by courtesy Center for Interest Measurement and Research, University of Minnesota.)

year-old female psychologist who is involved primarily in research and technical writing.[1] Figure 94 covers scores on the broadest scales, dealing with six General Occupational Themes, such as Realistic, Artistic, and Social. These themes are derived from the classification of interests developed by Holland (1973) and supported by extensive research both by Holland and by other independent investigators. Each theme characterizes not only a type of person but also the type of working environment that such a person would find most congenial. Scores on all parts of the inventory are expressed as standard scores ($M = 50$, $SD = 10$). For the General Occupational Themes, the normative group is the general reference sample consisting of 300 men and 300 women representative of all occupations covered by the inventory.[2] Numerical scores are reported with reference to the whole sample; interpretive phrases (e.g., high, moderately low) are based on the distribution of the same-sex subgroup.

The Administrative Indexes in Figure 94 are designed to detect carelessness and test-taking response sets. The two Special Scales are empirical scales developed against nonoccupational criteria. The Academic

---

[1] Data by courtesy of Center for Interest Measurement Research, University of Minnesota.

[2] For the detailed procedures followed to ensure representativeness, see D. P. Campbell (1971, Ch. 2; 1974, p. 49).

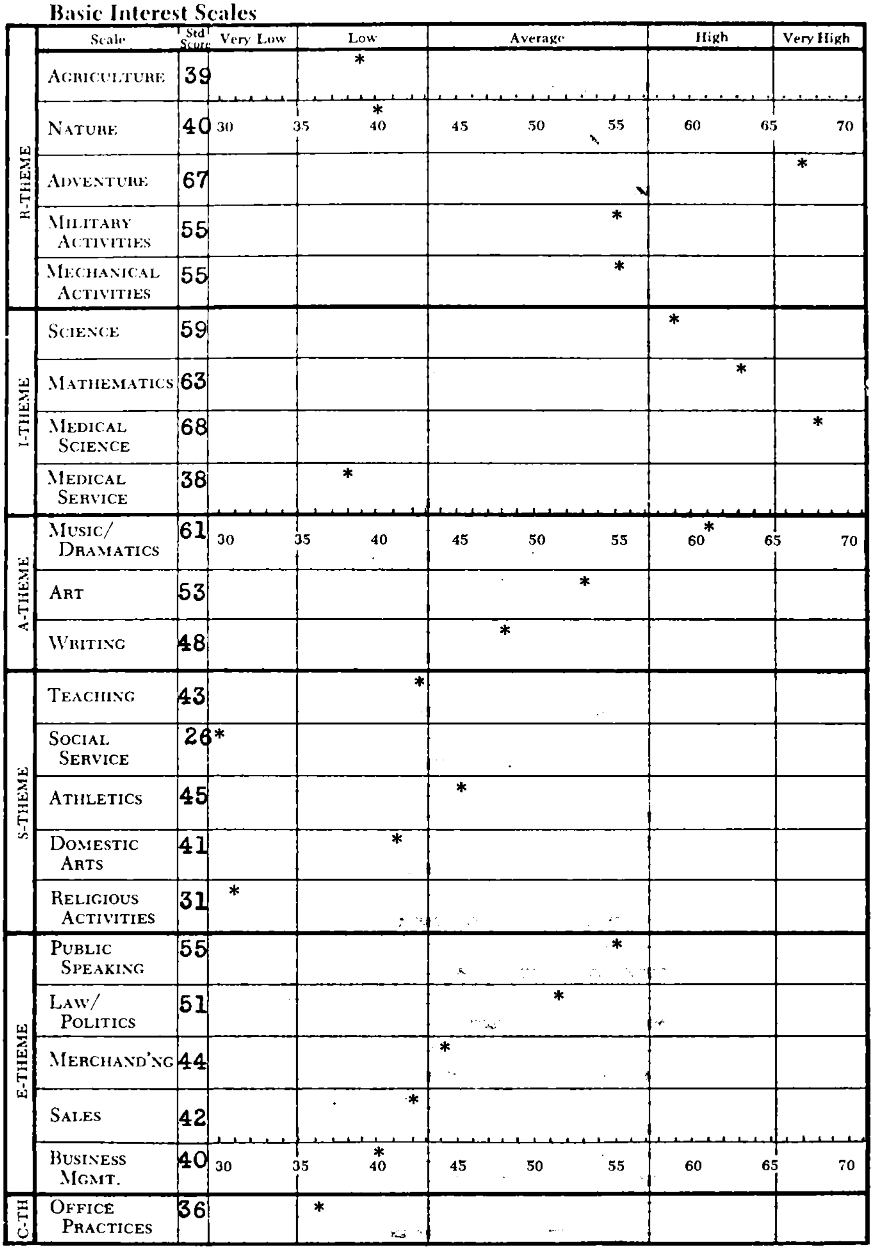
Orientation (AOR) scale was originally constructed by comparing item-response percentages of college students obtaining high grades with those of students obtaining low grades. It predicts chiefly the tendency to continue one's education through high school, college, and graduate school. The Introversion-Extroversion (IE) scale was developed from the responses of students identified as introverts and as extroverts on the MMPI. Scores reflect the person's interest in working alone or with people. Both scales have yielded large mean differences in the expected direction among groups engaged in different occupations.

Figure 95 shows the 23 Basic Interest Scales, classified under the six General Occupational Themes. These scales consist of clusters of substantially intercorrelated items. The Basic Interest Scales are more homogeneous in content than the Occupational Scales and can therefore help in understanding why an individual scores high on a particular Occupational Scale. The standard scores are again expressed in terms of the general reference sample; and the profile is plotted with reference to the same-sex subsample. For each scale, the open and shaded bars show the performance of the female and male subsamples, respectively.

The Occupational Scales, which constituted the main body of the SVIB since its inception, are illustrated in Figure 96. In the continuing research program on this inventory, new scales are constantly added and old scales updated with fresh criterion samples. The 1974 form provides 124 Occupational Scales, including 67 derived from male samples and 57 from female samples. Although the goal is to have male and female criterion samples for each occupation, by 1974 it had not proved feasible to gather data on both men and women for all the occupations on the form. Most of the samples had been tested in the 1960s; a few were tested in the 1970s and a few before 1960. Even today, it would be difficult to obtain a sufficiently large sample of men in some occupations or of women in others.

Regardless of sex of respondent, each inventory is scored for all Occupational Scales and standard scores are reported in terms of both male and female criterion groups for each occupation, when available. The profile, however, is plotted only against the same-sex norms. Occupational Scale scores show degree of similarity of the individual's responses to those of the criterion group for each occupation listed. Unlike the two broader types of scores, Occupational standard scores are derived from the appropriate occupational criterion groups, not from the general reference samples.

Reference to Figure 96 shows code letters preceding each occupational title. These letters indicate from one to three General Occupational Themes that predominate in that occupation. For example, the interests of psychologists are associated with Investigative (i.e., intellectual, scientific), Artistic, and Social themes. This classification was based

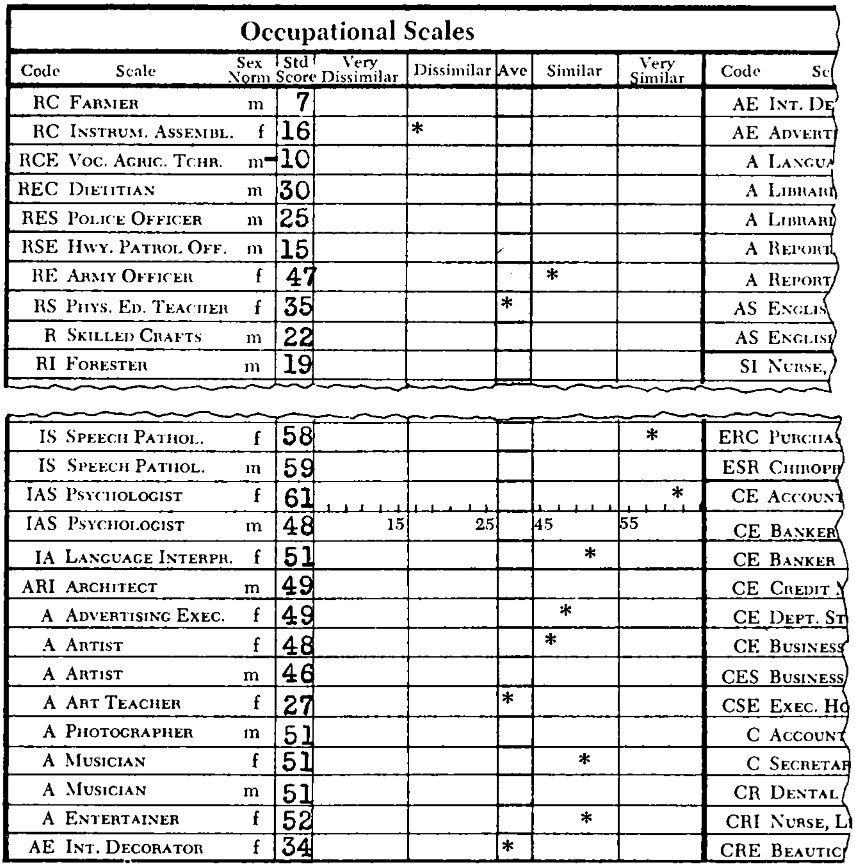


Basic Interest Scales

	Scale	Std Score	Very Low	Low	Average	High	Very High
R-THEME	Agriculture	39		*			
	Nature	40		*			
	Adventure	67					*
	Military Activities	55			*		
	Mechanical Activities	55			*		
I-THEME	Science	59				*	
	Mathematics	63				*	
	Medical Science	68					*
	Medical Service	38		*			
A-THEME	Music/ Dramatics	61				*	
	Art	53			*		
	Writing	48			*		
S-THEME	Teaching	43		*			
	Social Service	26	*				
	Athletics	45			*		
	Domestic Arts	41		*			
	Religious Activities	31	*				
E-THEME	Public Speaking	55			*		
	Law/ Politics	51			*		
	Merchand'ng	44			*		
	Sales	42		*			
	Business Mgmt.	40		*			
C-TH	Office Practices	36		*			

FIG. 95. Portion of SCII Profile, Showing Scores on 23 Basic Interest Scales, Classified According to General Occupational Themes. Shaded bars refer to responses of 300 men and open bars to responses of 300 women in General Reference Group; thick central portion covers range of middle 50%, thin portion range of middle 80%.

(Form reprinted with permission of the publisher from the *Manual for the Strong-Campbell Interest Inventory,* Form T325 of the STRONG VOCATIONAL INTEREST BLANK, by David P. Campbell [Stanford University Press, 1974], p. 18.)



Occupational Scales

Code	Scale	Sex Norm	Std Score	Very Dissimilar	Dissimilar	Ave	Similar	Very Similar	Code	Sc
RC	Farmer	m	7						AE	Int. De
RC	Instrum. Assembl.	f	16		*				AE	Advert
RCE	Voc. Agric. Tchr.	m-	10						A	Langua
REC	Dietitian	m	30						A	Librar
RES	Police Officer	m	25						A	Librar
RSE	Hwy. Patrol Off.	m	15						A	Report
RE	Army Officer	f	47				*		A	Report
RS	Phys. Ed. Teacher	f	35			*			AS	Englis
R	Skilled Crafts	m	22						AS	Englis
RI	Forester	m	19						SI	Nurse,
IS	Speech Pathol.	f	58					*	ERC	Purcha
IS	Speech Pathol.	m	59						ESR	Chirop
IAS	Psychologist	f	61					*	CE	Accoun
IAS	Psychologist	m	48	15	25		45	55	CE	Banker
IA	Language Interpr.	f	51				*		CE	Banker
ARI	Architect	m	49						CE	Credit
A	Advertising Exec.	f	49				*		CE	Dept. St
A	Artist	f	48				*		CE	Busines
A	Artist	m	46						CES	Business
A	Art Teacher	f	27			*			CSE	Exec. H
A	Photographer	m	51						C	Accoun
A	Musician	f	51				*		C	Secreta
A	Musician	m	51						CR	Dental
A	Entertainer	f	52				*		CRI	Nurse, L
AE	Int. Decorator	f	34			*			CRE	Beautic

Fig. 96. Portion of SCII Profile, Showing Scores on Some of the 124 Occupational Scales. Standard scores on original male and female Occupational Scales are reported for all examinees; profiles are plotted against like-sex groups. Letters preceding scale titles refer to General Occupational Themes statistically related to each Occupational Scale.

(Form reprinted with permission of the publisher from the *Manual for the Strong-Campbell Interest Inventory,* Form T325 of the Strong Vocational Interest Blank, by David P. Campbell [Stanford University Press, 1974], p. 18.)

principally on the mean scores on the General Occupational Theme scales obtained by each occupational criterion sample, and on the correlations of the Occupational Scales with each of the Theme Scales.

Items for each Occupational Scale were selected and weighted on the basis of *differences* in item-response percentages between the occupational criterion sample and the reference sample. In the Farmer Scale, for example, a +1 weight indicates that the response occurs more frequently, and a −1 that it occurs less frequently, among farmers than among men-in-general. Responses that fail to differentiate farmers from men-in-general do not appear in the Farmer Scale, regardless of how frequently they were chosen by farmers. An individual's total raw score on each Occupational Scale is simply the algebraic sum of his plus and minus weights. The raw scores are converted to standard scores in terms of the distribution of scores in each occupational criterion group. These criterion groups, usually containing from 100 to 500 cases, included persons between the ages of 25 and 55, employed in the given occupation for at least three years, who reported satisfaction with their work and met certain minimal standards of successful achievement.

The SVIB-SCII has been subjected to a continuing research program that has yielded extensive data about its reliability and validity (see D. P. Campbell, 1971, 1974). Median retest reliabilities over a 30-day period for General Occupational Themes, Basic Interest Scales, and Occupational Scales are all in the high .80's. Long-term stability of the Occupational Scales for periods ranging from one year to over 20 years is also high, the correlations falling mostly in the .60's and .70's. Another type of analysis concerns the stability of mean occupational profiles over time (D. P. Campbell, 1971, Ch. 9). For this purpose, the SVIB was administered to different samples of men holding the same jobs in the same organizations on two occasions separated by an interval of 30 years or more. This procedure was followed with four occupational groups: ministers, bankers, school superintendents, and corporation presidents. In general, the mean occupational profiles obtained over this interval were remarkably similar, suggesting that the scales developed on the original criterion groups are still applicable today. Similar results have been obtained with the Basic Interest Scales administered to different samples from the same occupation over a 30-year span.

Concurrent validity is indicated by the degree of differentiation among different occupational samples, and between occupational samples and reference samples. Predictive validity has been checked in several samples over long intervals. The evidence indicates substantial correspondence between the initial occupational profile and the occupation eventually pursued. A specific example is provided by a 40-year follow-up of a sample of psychologists, whose professional careers revealed some suggestive relations between the flatness versus distinctness of their original profile and such occurrences as frequency of job changes and shifting from teaching or research to administrative or applied work (Vinitsky, 1973). Another investigation found strikingly high cross-cultural profile similarities among samples of psychologists tested in nine Western nations (Lonner & Adams, 1972).

With regard to construct validity, the relations of the Occupational Scales to the General Occupational Themes, as well as the interrelations of the themes among themselves, are of special relevance. Figure 97

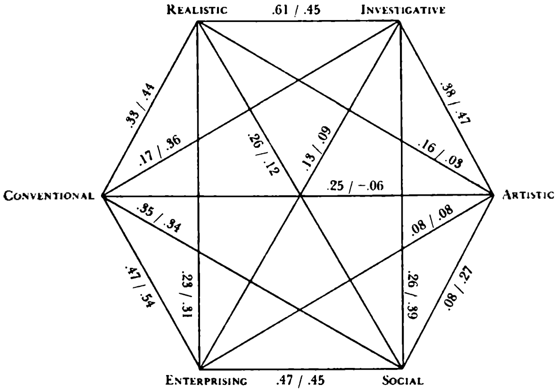


FIG. 97. Holland's Hexagonal Model of General Occupational Themes. For each pair of themes, first correlation is based on female sample, second correlation on male sample.

(Reprinted with permission of the publisher from the *Manual for the Strong-Campbell Interest Inventory*, Form T325 of the STRONG VOCATIONAL INTEREST BLANK, by David P. Campbell [Stanford University Press, 1974], p. 34.)

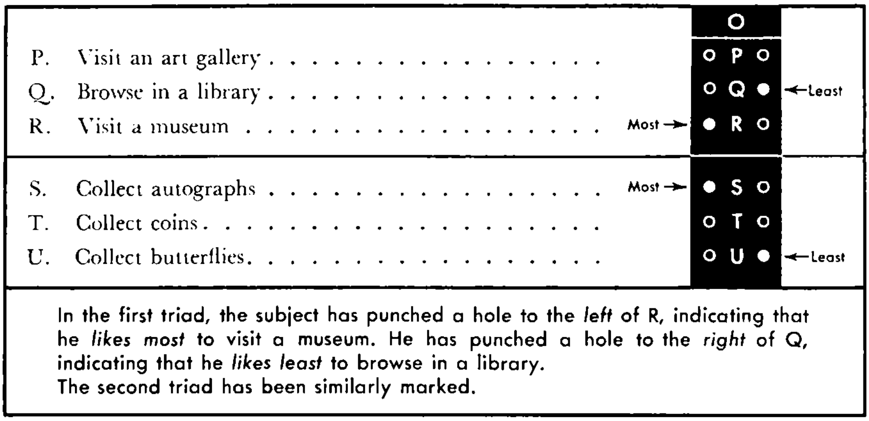
shows the six themes at the corners of the hexagonal model developed by Holland (1966, 1973). It will be noted that the highest correlations, in both female and male samples, were obtained between theme scales occupying adjacent positions along the perimeter of the hexagon. The lowest correlations were found between scales at opposite ends of diagonals. Among women, for example, the Realistic Scale correlates .61 with the Investigative Scale, but only .16 with the Artistic and .26 with the Social Scale. Similarly, if the Occupational Scales are plotted on the hexagon, most of them follow the expected order along the perimeter. For example, Farmer, coded RC, falls between Realistic and Conventional; Psychologist, coded IAS, falls along the Investigative, Artistic, Social continuum. Usually, occupations that score high on one theme score low on its direct opposite (e.g., Artistic and Conventional). When an Occupational Scale shows substantial correlations with themes at opposite points of the hexagon, it often includes heterogeneous subgroups with recognizably different occupational functions.

To round out this discussion of the SVIB, it should be noted that a similar interest inventory has been developed for various skilled and semiskilled occupations (D. P. Campbell, 1966; K. E. Clark, 1961; Clark & Campbell, 1965). Modeled after the SVIB but utilizing tradesmen-in-general as a reference group, the Minnesota Vocational Interest Inventory (MVII) provides 21 scales for such occupations as baker, plumber, and radio-TV repairman. There are also nine area scales showing the examinee's liking for work activities common to several occupations, such as mechanical, food service, and electronics. These scales will be recognized as analogous to the Basic Interest Scales in the SCII. The items in the Minnesota Vocational Interest Inventory are more heavily oriented toward vocational activities than are those of the SVIB, and they are all presented in a forced-choice triad format. Differentiation among occupations is not as sharp with this inventory as with the SVIB. It is likely that this difference between the two instruments stems from the nature of the occupations themselves. There is extensive evidence to show that at higher levels of the occupational hierarchy (professional, managerial, etc.) job satisfaction is derived chiefly from intrinsic liking for the work, while at the lower levels there is increasing reliance on such extrinsic factors as pay, security, social contacts, and recognition as a person (Darley & Hagenah, 1955). The measurement of vocational interest patterns may thus become less relevant as we go down the occupational hierarchy.

THE KUDER INTEREST INVENTORIES. The various forms, versions, and editions of the Kuder interest inventories may be regarded as a family of related instruments which approach the measurement of interests from different angles and are designed for somewhat different purposes. The earliest and best known of these inventories is the Kuder Vocational Preference Record. Developed later than the SVIB, this test followed a different approach in the selection and scoring of items. Its major purpose was to indicate relative interest in a small number of broad areas, rather than in specific occupations. The items were originally formulated and tentatively grouped on the basis of content validity. This was followed by extensive item analyses on high school and adult groups. The object of such item analyses was the development of item groups showing high internal consistency and low correlations with other groups. This aim was reasonably well fulfilled for most of the scales.

The items in the Kuder inventories are of the forced-choice triad type. For each of three activities listed, the respondent indicates which he would like most and which he would like least. Two sample triads are illustrated in Figure 98. The Kuder Vocational Preference Record provides 10 interest scales plus a Verification scale for detecting carelessness, misunderstanding, and the choice of socially desirable but unlikely answers. The interest scales include: Outdoor (agricultural, naturalistic), Mechanical, Computational, Scientific, Persuasive, Artistic, Literary, Musical, Social Service, and Clerical. The similarities between these interest scales and the six themes later identified in Holland's research are clearly apparent. Separate sex norms are available for high school, college, and adult groups. Total scores in the 10 interest areas are plotted on a normalized percentile chart.

The reliabilities of the Kuder scales, as determined by the Kuder-



			○	
P.	Visit an art gallery		○ P ○	
Q.	Browse in a library		○ Q ●	←Least
R.	Visit a museum	Most→	● R ○	
S.	Collect autographs	Most→	● S ○	
T.	Collect coins		○ T ○	
U.	Collect butterflies		○ U ●	←Least

In the first triad, the subject has punched a hole to the *left* of R, indicating that he *likes most* to visit a museum. He has punched a hole to the *right* of Q, indicating that he *likes least* to browse in a library.
The second triad has been similarly marked.

FIG. 98. Sample Items from Kuder Vocational Preference Record.
(Reproduced by permission of Science Research Associates, Inc.)

Richardson technique, cluster around .90. Stability over intervals of about a year or less also appears to be satisfactory. Little information is available regarding stability over longer periods. There is some evidence to suggest that, especially in the case of high school students, shifts in high and low interest areas are relatively frequent when retests are several years apart (Herzberg & Bouton, 1954; Mallinson & Crumrine, 1952; Reid, 1951; Rosenberg, 1953). Studies on the simulation of interest scores have shown that faking is possible to some extent on both the Kuder and the Strong; but visibility is somewhat greater on the Kuder, owing to the more obvious nature of its items (Durnall, 1954; Longstaff, 1948).

The manual for the Kuder provides an extensive list of occupations, grouped according to their major interest area or pair of interest areas. For example, radio operator is classified under Mechanical; landscape architect, under Outdoor-Artistic. This is an a priori listing in terms of logical or content analysis. In addition, empirically established occupational profiles are given for 41 homogeneous occupational families. The data employed in deriving these occupational profiles were contributed largely by test users. Consequently, several of the groups are small and their representativeness or comparability is questionable. Efforts have also been made to develop regression equations for finding total scores for specific occupations or other criterion groups (Mugaas & Hester, 1952).

Validation research with this inventory has been conducted chiefly against a criterion of job satisfaction.[3] In the most comprehensive longitudinal study, 1,164 students who had taken the Kuder Vocational Preference Record in high school were given a job satisfaction questionnaire seven to ten years later. At that time, 728 were engaged in work classified as "consistent" with their original interest patterns and 436 were engaged in work classified as "inconsistent" with these patterns. The percentage of satisfied workers was 62 in the "consistent" group but only 34 in the "inconsistent" group; the percentage of dissatisfied workers, on the other hand, was 8 in the "consistent" group and 25 in the "inconsistent" group. Similar relationships were found in several studies of persons employed in specific fields, such as clerks, industrial workers, vocational rehabilitation counselors, and accountants. Among accountants, for example, the satisfied group scored significantly higher than the dissatisfied group on the Computation and Clerical interest scales, and significantly lower on the Outdoor, Scientific, and Artistic interest scales.

---

[3] References to this research can be found in Kuder (1964), pp. 39–40.

More recently, the Kuder General Interest Survey (KGIS) has been developed as a revision and downward extension of the Kuder Vocational Preference Record. Designed for grades 6 to 12, the General Interest Survey employs simpler language and easier vocabulary than the earlier form, requiring only a sixth-grade reading level. Percentile norms are reported for a national stratified sample of over 10,000 boys and girls and for small adult samples. In many situations, the KGIS has replaced the earlier Kuder Vocational Preference Record. Correlations between scores on the two instruments are nearly as high as their reliability coefficients. Most of the research has been conducted with the earlier form. Despite the similarity between the two forms, however, more data regarding the KGIS itself would be desirable, particularly with younger populations.

Still another version, the Kuder Occupational Interest Survey (KOIS), has been developed through criterion keying procedures similar to those followed with the SVIB (Kuder, 1966a, 1966b, 1970). Unlike the SVIB and an earlier Kuder occupational form, however the KOIS does not employ a general reference group. Instead, the individual's score on each occupational scale is expressed as a correlation between his interest pattern and the interest pattern of the given occupational group.[4] The KOIS cannot be hand-scored; answer sheets are returned to the publisher for computer scoring. Scores are currently available for 77 occupational and 29 college-major scales derived from male criterion groups; and for 29 occupational and 19 college-major scales derived from female criterion groups. Scores on all scales are reported for both male and female respondents. The occupations covered by this inventory vary widely in level, ranging from baker and truck driver to chemist and lawyer. The elimination of a reference group permits this broad coverage within a single instrument.

---

[4] The correlation employed is the lambda coefficient developed by Clemans (1958). This is essentially a point-biserial correlation adjusted for differences in homogeneity of different criterion groups. The dichotomous variable consists of the marked versus unmarked responses on the individual's answer sheet; the continuous variable is the proportion of persons in the criterion group marking each response.

Through intensive statistical analyses of the scores of 3,000 subjects (100 in each of 30 core groups representative of the occupations and college-major fields covered by the inventory), Kuder has demonstrated that better differentiation between occupational groups can be achieved with the KOIS scoring system than with the occupational scales derived through the use of a general reference group. The same 30 scales have been employed in extensive analyses of retest reliability, intercorrelation of scores on different scales, and other technical aspects of the inventory. In all these analyses, the KOIS appears to be highly satisfactory. Research on the other occupational scales, as well as validation against external criteria, remains to be done. This instrument presents some major innovations in test construction. Its ultimate contribution can be assessed only after further research and application in practical contexts.

SOME METHODOLOGICAL QUESTIONS. The development of interest inventories highlights several important problems in test construction. A major controversial question pertains to the use of ipsative scores. Some of the implications of these scores have already been discussed in Chapter 17, in connection with the EPPS. It will be recalled that ipsative scores permit comparisons of the relative strength of characteristics within the individual, while normative scores permit external comparisons with a normative sample. When this distinction is overlooked or when both types of scores are combined, confusion is likely to result in both score interpretation and statistical analysis of data.

Interest inventories have traditionally relied on ipsative scores. This is true of both the SVIB and the Kuder inventories. Although the ipsative nature of the Kuder scores is more obvious, because of the triad item form in which different interests are contrasted, the SVIB scores also have ipsative properties. This is so not only because one section of the SVIB requires a choice between two response alternatives, but also because individual responses are scored on more than one scale. Hence, scores on different scales are not experimentally independent. For example, insofar as the same response may be positively weighted on one scale and negatively weighted on another, it is impossible for an individual to obtain high scores or low scores on all scales.

The controversy regarding the use of ipsative scores is far from resolved. While several critics have objected to the use of ipsative scores, particularly when combined with normative interpretations, Kuder has defended the practical meaningfulness of such scores. Thus, he argues that, if a student obtains a percentile score of 90 on the Mechanical scale of the General Interest Survey, it means that, "when faced with a complex series of choices typical of real-life situations, he chose mechanical activities more frequently than 90 percent of his contemporaries" (Kuder, 1964, p. 6). Others have argued against such a combination of ipsative and normative approaches, because of the difficulties of interpreting these scores and the frequency with which they have been analyzed by inapplicable statistical procedures (Bauernfeind, 1962; M. R. Katz, 1962). It has also been objected that ipsative interest scores partial out any individual differences in level of interest or enthusiasm, although such differences may be germane to many practical decisions. While the controversy remains unsettled, the test user should be aware of the ipsative nature of interest inventory scores and their interpretive implications.

A second methodological question pertains to the alternative approaches to interest measurement represented by criterion-keyed occupational scales (as in the original SVIB and in the KOIS) and by assessment of broad interest areas (as in the Kuder General Interest Survey). It should be noted that, even with the latter approach, an individual's interest in specific occupations can be appraised by such procedures as occupational coding of interest profiles or the computation of regression equations for specific occupations. Moreover, broad interest areas and occupational scales can be included in the same instrument, as in the SVIB-SCII.

When occupational scales are employed, a further question concerns the use of reference groups. Such groups have been utilized in the development of the SVIB-SCII, the Minnesota Vocational Interest Inventory, and the earlier form of Kuder's occupational interest inventory. More recently, Kuder has argued convincingly for the derivation of occupational scores without a reference group (Kuder, 1963, 1966b, 1970). It is certainly theoretically true that the direct comparison of one occupation with another is likely to elicit sharper differentiation between them than is obtained when each occupation is compared with a miscellaneous group. The elimination of a reference group also sweeps away several persistent problems regarding the selection of suitable reference groups, in terms of socioeconomic level, occupational diversity, and other demographic variables. On the other hand, the scoring procedure followed in the KOIS, as a result of the elimination of the reference group, is so time-consuming as to preclude hand scoring. However, this may not constitute a practical limitation in our society today. In this connection, it is interesting to note that the latest forms of both the Kuder Occupational Interest Inventory and the SVIB-SCII require computer scoring.

THE PLACE OF INTERESTS IN PERSONALITY THEORY. The measurement of interests began as a relatively narrow and specialized development in the study of personality. Early interest inventories were oriented chiefly towards the prediction of the individual's eventual acceptance or rejection of particular job functions. From these modest beginnings, interest tests have gradually come to play a significant role in the formulation of personality theory. The impetus for these developments has come from several sources.

Factorial analyses of interest items, as illustrated by the work of Guilford and his associates (Guilford, 1959; Guilford et al., 1954), first suggested the interconnections of interests with other important dimensions of personality. Scores on interest inventories, such as the Strong and Kuder, have been included in correlational or factor-analytic studies with other self-report inventories, including the MMPI, Personality Research Form, Adjective Check List, Study of Values, and Rokeach Dogmatism Scale (Darley & Hagenah, 1955; Dunteman & Bailey, 1967; Hansen & Johansson, 1974; Scott & Day, 1972; Siess & Jackson, 1967). All these studies have found substantial relationships between several broad emotional and attitudinal variables, on the one hand, and occupational interests, on the other.

Comparative investigations of several behaviorally defined groups have likewise revealed significant correspondences between vocational interests and other aspects of personality. For example, a series of studies on high school and college students, industrial workers, and neuropsychiatric patients found accident-proneness to be significantly related to scores on some of the SVIB Occupational Scales (Kunce, 1967). Similarly, research with several psychiatric groups, including neurotics and alcoholics, revealed a predominance of certain vocational interest patterns in each group (Carnes, 1964; Drasgow & Carkhuff, 1964; Force, 1958; Patterson, 1957).

Detailed personality evaluations of persons receiving high scores on different occupational interest scales are provided by a study of 100 Air Force officers conducted at the University of California (see Darley & Hagenah, 1955, pp. 128–129). Each subject was given a battery of tests, including the SVIB, and also underwent an intensive assessment program through interviews and other observational techniques. On the basis of the available information, subjects were described by eight clinical psychologists in terms of 76 designated personality variables. Correlations of these trait ratings with each of the SVIB occupational keys revealed a number of statistically significant relations. For illustrative purposes, the personality descriptions associated with high scores on two Occupational Scales are summarized below:

High Scorers on Mathematician Key: Self-abasing, concerned with philosophical problems, introspective, lacking in social poise, lacking confidence in own ability, sympathetic, reacts poorly to stress, not persuasive in personal contacts, not an effective leader, not ostentatious, not aggressive or socially ascendant.

High Scorers on Real Estate Salesman Key: Self-indulgent, guileful, cynical, opportunistic, aggressive, persuasive, ostentatious, may arouse hostility in others, not sympathetic, not concerned with philosophical problems, not lacking confidence in own ability, not self-abasing.

It is now widely recognized that the choice of an occupation often reflects the individual's basic emotional needs and that occupational adjustment is a major aspect of general life adjustment (Bordin, 1943; Darley & Hagenah, 1955; Holland, 1966, 1973; Roe, 1956; Super, 1957). There are many different ways of dealing with interpersonal relations and other life problems. No one way is universally better than others. When he chooses a vocation, each individual is to some extent selecting those adjustment techniques, life patterns, and roles most congenial to himself. The measurement of vocational interests—and more specifically the identification of those occupational groups whose interests and attitudes the individual shares most closely—thus becomes a focal point in the understanding of different personalities. Direct studies of the characteristics of persons in different occupations have been contributing to a growing fund of factual material for implementing this approach.

Such an approach to personality assessment is illustrated by the work of Holland (1966, 1973), whose hexagonal model of occupational themes has been incorporated in the SVIB-SCII. Holland aligns himself clearly with those who regard occupational preferences as the choice of a way of life—a choice that reflects the individual's self-concept and his major personality characteristics. Each of Holland's occupational themes corresponds to a "type" or cluster of personal attributes. A given individual may be described in terms of one or more predominant types. The occupational themes also correspond to model environments in terms of which different occupational environments may be characterized. These environments comprise not only physical features and work demands but also the kinds of persons with whom the individual works (co-workers, supervisors, customers, patients, students, etc.). According to Holland, individuals seek environments that are congruent with their personality types; and such congruence enhances work satisfaction, job stability, and achievement.

From a different angle, Tyler (1959, 1974; Tyler, Sundberg, Rohila, & Greene, 1968) regards the study of interests as a way of identifying the choices that the individual makes at various life stages. These choices are both a reflection of the kind of person he is and a forecast of what he is likely to become. As each choice is made—choice of friends, recreations, courses, jobs, and the like—the individual's subsequent experiences are thereby channeled into certain paths. Alternative developmental routes are eliminated at each choice point. Translating the predictive validity of the SVIB into these terms, Tyler (1959, p. 78) wrote:

> [A high Occupational Scale score] signifies that the person's characteristic pattern of acceptance and rejection of life's varied possibilities is like the choice pattern characteristic of persons in a certain occupation. What we should expect then to be able to predict from such a score is . . . the way he will make his choices at later junctures of his life. This makes sense of the high degree of validity Strong's recent studies have shown for the test.

The measurement of interests promises to be a lively field of test development and research in the years ahead. Having proved so far to be among the most successful tests outside the aptitude domain, interest inventories are now well on the way to attaining theoretical respectability. At the same time, we can anticipate that the interpretation of interest profiles will show further advances in depth and sophistication.

OPINION AND ATTITUDE MEASUREMENT

NATURE OF INSTRUMENTS. An attitude is often defined as a tendency to react favorably or unfavorably toward a designated class of stimuli, such as a national or racial group, a custom, or an institution. It is evident that, when so defined, attitudes cannot be directly observed, but must be inferred from overt behavior, both verbal and nonverbal. In more objective terms, the concept of attitude may be said to connote response consistency with regard to certain categories of stimuli. In actual practice, the term "attitude" has been most frequently associated with social stimuli and with emotionally toned responses.

Opinion is sometimes differentiated from attitude, but the proposed distinctions are neither consistent nor logically defensible. More often the two terms are used interchangeably, and they will be so employed in this discussion. With regard to assessment methodology, however, opinion surveys are traditionally distinguished from attitude scales. *Opinion surveys* are characteristically concerned with replies to specific questions, which need not be related. The answers to such questions are kept separate rather than being combined into a total score. An employee opinion survey, for example, might include questions about work schedules, rate of pay, fringe benefits, company cafeteria, and relation to supervisors; each of these items is included because of its intrinsic relevance to the improvement of employee relations. The replies to each question are separately tabulated in the effort to identify sources of employee satisfaction and dissatisfaction.

Attitude scales, on the other hand, typically yield a total score indicating the direction and intensity of the individual's attitude toward a company, group of people, policy, or other stimulus category. In the construction of an attitude scale, the different questions are designed to measure a single attitude or unidimensional variable, and some objective procedures are usually followed in the effort to approach this goal. An employee attitude scale, for example, yields a single score showing the individual's degree of job satisfaction or overall attitude toward the company.

APPLICATIONS. Both opinion surveys and attitude scales have been widely used for a variety of purposes. *Public opinion polling* is a familiar application. Besides the election forecasts undertaken by commercial pollsters, nationwide polls are conducted regularly on many social, political, economic, and international questions of general interest. The large-scale nature of these polls and the rapidity with which answers are required have encouraged the utilization of survey, rather than attitude-scale, techniques. But attitude scales have also found a place in public opinion research. *Market research* has much in common with public opinion studies. Because of similar practical demands, survey techniques have served as the principal tool in both fields. The object of market research is to investigate consumer needs and reactions in reference to products, services, or advertisements. The resulting information may be used for such purposes as choosing the most effective advertisement for a specific article, improving a type of service, preparing a new model, or designing a new product to meet consumer specifications.

Another major field of application is to be found in the measurement of *employee attitudes and morale* through both opinion surveys and attitude scales. Such techniques are also employed to assess the effectiveness of *education and training.* Attitude scales, for example, may provide an index for evaluating different instructional procedures. Or they may be utilized in measuring the changes in student attitudes toward literature, art, different racial and cultural groups, or social and economic problems following a given educational program. One of the most extensive applications of attitude measurement is to be found in *research in social psychology.* Practically every textbook on social psychology contains sections on attitudes and their measurement. Among the many problems investigated through attitude measurement may be mentioned group differences in attitudes, the role of attitudes in intergroup relations, background factors associated with the development of attitudes, the interrelations of attitudes (including factor analyses), trends and temporal shifts in attitudes, and the experimental alteration of attitudes through interpolated experiences.

METHODOLOGICAL PROBLEMS. The measurement of attitudes is both difficult and controversial.[5] Whether verbally expressed opinions can be regarded as indicators of "real" attitudes has frequently been questioned. In part, this problem concerns the relationship between verbal and nonverbal overt behavior. In other words, does the individual suit his actions to his words—or to his attitude scale score? Discrepancies between verbally expressed attitudes and overt behavior have been noted in a number of studies. It has been further pointed out that even observations of overt behavior may not always provide an accurate index of attitude. For example, an individual may both profess strong religious beliefs and attend church regularly, not because of his religious convictions but as a means of gaining social acceptance in his community. Such a possibility raises the further question of the relationship between "public" and "private" attitudes. How do the individual's publicly expressed attitudes compare with the opinions he voices in conversation with his family or close friends? Public opinion surveys are usually "public" in more than one sense. They represent a verbal expression of attitudes by the public and to the public. The question of situational specificity and of the generalizability of attitude measures must also be considered (LaFave, Haddad, & Marshall, 1974; Mischel, 1968; D. Peterson, 1968).

Kelman (1974) gives a searching analysis of the relation between attitudes and action, supported by a systematic survey of research results. His conclusion is that attitudes provide a useful personality construct, if viewed as an integral part of action. Although not an index of action, verbally reported attitudes serve as determinants, components, and consequents of action. For example, attitudes may contribute toward action changes; and actions may contribute toward attitude changes. As Kelman (1974, p. 316) expresses it, "Attitudes and action are linked in a continuing reciprocal process, each generating the other in an endless chain."

From another viewpoint, the relationships between "what a person says" and "what he does," between publicly and privately expressed attitudes, and between attitudes manifested in different situations may be regarded as special instances of validity. Attitude scales and opinion polls may be validated against a number of criteria, such as membership in contrasted groups, ratings by close acquaintances, and biographical data secured through intensive interviews or case studies. Because of the practical difficulties in obtaining such criterion data, however, investigators have frequently relied on the familiar makeshifts involving validation by internal consistency or by correlation with another attitude scale. Other investigators have resorted to a superficial kind of content validation based on the examination and classification of questions according to the topics covered. For opinion polling, validation is rarely attempted at all.

---

[5] For a comprehensive introduction to relevant literature, see Fishbein (1967), Summers (1970).

Data on reliability are also meager, especially with reference to opinion surveys. Yet in the case of survey techniques, with their reliance on single questions, reliability is most likely to be suspect. Attitude measurement also presents a number of other methodological problems. These problems are not fundamentally different from those encountered in the construction and administration of other types of psychological tests, but they are accentuated in the measurement of attitudes. The major difficulties center around the proper formulation of questions so as to avoid ambiguity, suggestion, and other sources of error; adequate sampling of the population in terms of size and representativeness; and control of conditions under which the survey is conducted, such an ensuring anonymity of replies and reducing the influence of interviewer characteristics.

All of the problems considered above—including validity, reliability, questionnaire construction, sampling, and administration—are encountered in varying degrees in both opinion surveys and attitude measurement. From a technical viewpoint, attitude scales are clearly superior to opinion surveys. In their construction and use, attitude scales are more nearly similar to psychological tests. Although by their very nature most opinion surveys and many attitude scales must be custom-made to meet specific needs, at least some attitude scales have been developed for general use. In the following section, we shall examine in more detail the procedures employed in the development and application of attitude scales.

ATTITUDE SCALES

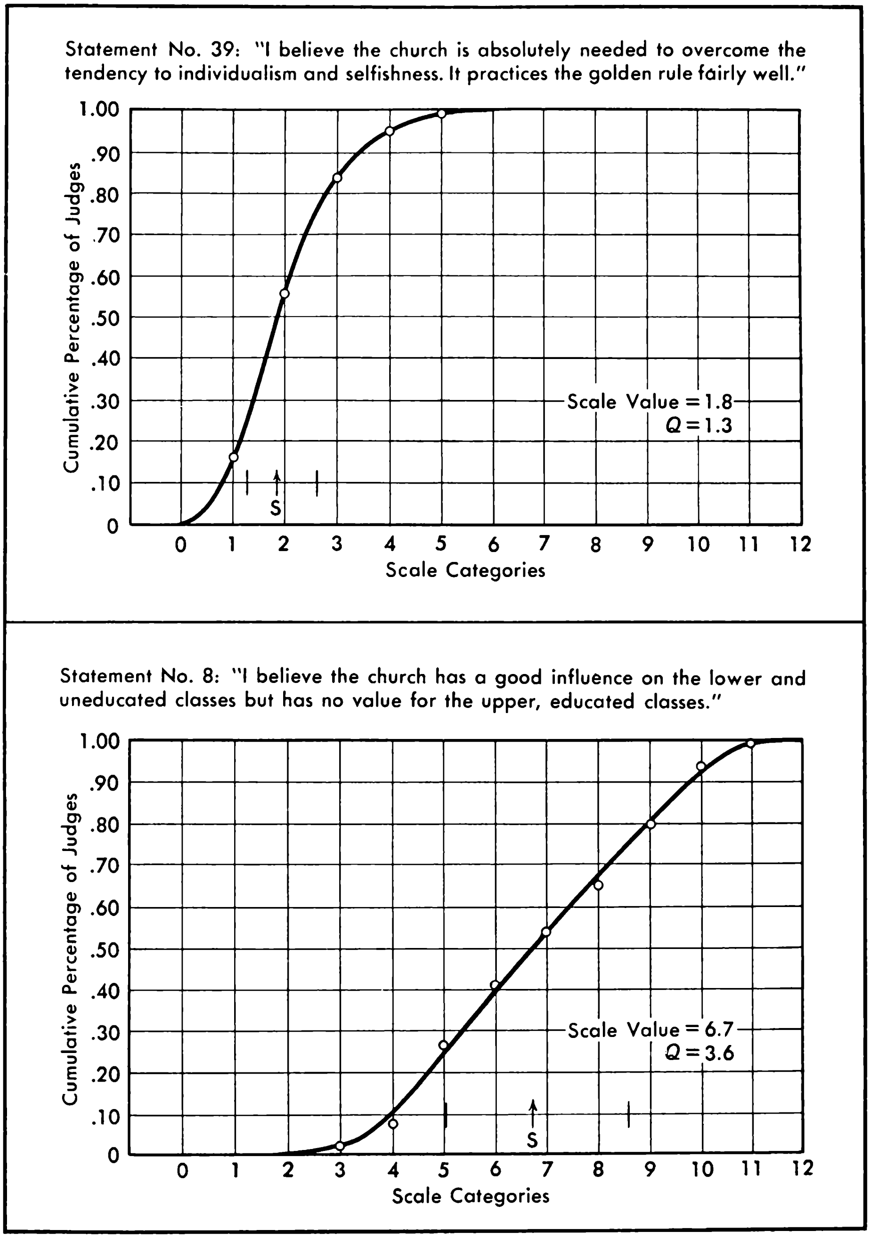
THURSTONE-TYPE SCALES. Attitude scales are designed to provide a quantitative measure of the individual's relative position along a unidimensional attitude continuum. Special procedures have been devised in an attempt to achieve comparability of scores from scale to scale, equality of distances between scale units, and unidimensionality or homogeneity of items. Thurstone's adaptation of psychophysical methods to the quantification of judgment data represented an important milestone in attitude scale construction (Thurstone, 1959; Thurstone & Chave, 1929). By these procedures, Thurstone and his co-workers prepared about 30 scales for measuring attitudes toward war, capital punishment, the church, patriotism, censorship, and many other institutions, practices, issues, and racial or national groups.

The construction of the Thurstone-type scales may be illustrated by considering the Scale for Measuring Attitude toward the Church, since the development of this scale was fully reported in published form (Thurstone & Chave, 1929). Essentially the same procedure was followed in preparing all other scales in the series. The first step was to gather a large number of statements regarding the church. These statements were obtained principally by asking several groups of people to write out their opinions about the church. The list was supplemented with statements taken from current literature. A search was made for expressions of opinion ranging from extremely favorable, through neutral, to extremely unfavorable. From the material thus collected, a list of 130 carefully edited, short statements was drawn up.

These statements, each mimeographed on a separate slip, were given to each of 300 judges for sorting into 11 piles, from A to K. The judges were instructed to put in category A those statements they believed expressed the highest appreciation of the value of the church, in category F those expressing a neutral position, and in category K those expressing the strongest depreciation of the church. In the intervening piles, statements were to be arranged in accordance with the degree of appreciation or depreciation they expressed. This sorting procedure has been described as the method of "equal-appearing intervals," although the judges were not actually told that the intervals between piles were to appear equal. It should also be noted that the judges were not asked to indicate their own attitudes toward the church, but were requested only to classify the statements.

The percentage of judges who placed each statement in the different categories constituted the basic data for computing the scale values of the statements. The method is illustrated in Figure 99. On the baseline of the cumulative frequency graph are the numbers 1 to 11, corresponding to categories A to K, which are treated as equally spaced points on the scale. The vertical axis shows the percentage of judges placing the statement in or below each category. The 50th percentile, or median position assigned by the judges to the statement, can be read directly from the graph. This median position is taken as the *scale value* of the statement. It will be noted that for one of the statements illustrated (No. 39), which is quite favorable to the church, the scale value is 1.8. The other statement (No. 8) has a scale value of 6.7. Once the scale values were computed for all statements, the next step was to select those statements whose scale values were equally spaced along the attitude continuum.

Besides the scale value, or median position, of each statement, the graphs also show the variability or spread of positions assigned to it by



Statement No. 39: "I believe the church is absolutely needed to overcome the tendency to individualism and selfishness. It practices the golden rule fairly well."

Statement No. 8: "I believe the church has a good influence on the lower and uneducated classes but has no value for the upper, educated classes."

FIG. 99. Determination of Scale Values for Thurstone-Type Attitude Scale. (Adapted from Thurstone & Chave, 1929, pp. 37, 39; reproduced by permission of University of Chicago Press.)

the different judges. The index of variability employed for this purpose (*Q*) is simply the distance between the 25th and 75th percentile points. Reference to Figure 99 shows that the judges agreed closely in the placement of statement 39, but varied widely in classifying statement 8. This difference is reflected in the *Q*'s of 1.3 and 3.6, respectively, which were obtained for the two statements. This measure of variability was taken as an index of the *ambiguity* of statements. Thus, ambiguous or double-barreled statements, which are variously interpreted by different judges, would tend to be less consistently classified. Accordingly, statements yielding high *Q*'s were eliminated from the final scale.

The statements were also checked for *irrelevance*. This was accomplished by presenting the 130 statements to subjects with the instructions to mark those statements with which they agreed. The responses were then analyzed statistically to determine their internal consistency. Statements that failed to meet the criterion of internal consistency were excluded as being irrelevant to the variable under consideration. The final scales thus comprise items that proved to be relatively unambiguous, relevant, and evenly distributed over the range of scale values.

In taking any Thurstone-type attitude scale, the respondent marks all statements with which he agrees. His score is the median scale value of the statements he has endorsed. The Thurstone technique has been applied to the development of attitude scales for many purposes. Thurstone-type scales have been constructed, for example, to measure the attitude of employees toward their company. A few statements from one of these scales, with their corresponding scale values, are shown below (Uhrbrock, 1934):

I think this company treats its employees better than any other company does	10.4
If I had to do it over again I'd still work for this company	9.5
The workers put as much over on the company as the company puts over on them	5.1
You've got to have "pull" with certain people around here to get ahead	2.1
An honest man fails in this company	0.8

A difficulty inherent in the development of Thurstone-type scales pertains to the possible effects of the judges' own attitudes on their classification of the statements. Thurstone recognized this problem, stating that "if the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it," and adding "until experimental evidence may be forthcoming, we shall make the assumption that the scale values of the statements are independent of the attitude distribution of the readers who sort the statements (Thurstone & Chave, 1929, p. 92). A number of early studies corroborated Thurstone's original assumption, insofar as scale values did not differ appreciably when redetermined on groups known to differ in their own attitudes.

Later investigations, however, found that under certain conditions scale values are significantly affected by judges' attitudes (Farnsworth, 1943; Fishman & Lorge, 1959; Hovland & Sherif, 1952; Sherif, Sherif, & Nebergall, 1965). Thus, large and significant shifts in the scale values of statements about war occurred from 1930 to 1940. Similar differences were obtained with the scale on attitude toward blacks, when items were rescaled on several black and white groups. The constancy of scale values reported by the earlier studies appeared to have resulted in part from the failure to include subjects with a sufficiently wide range of attitudes. Moreover, it was a common practice in these studies, as in Thurstone's original work, to discard the records of judges who placed 30 or more statements in a single pile. Such a massing of statements was regarded as evidence of careless sorting. It has subsequently been demonstrated, however, that these disproportionate groupings are more likely to occur among individuals whose own views are extreme.

In general, intergroup differences in item placement are reduced when ambiguous items with large inter-judge variability (Q) are eliminated and when only items separated by large scale differences are chosen. On the other hand, if judges are presented with ambiguous and relatively neutral items and if the conditions of judgment are not highly controlled —as when judges are allowed to determine how many categories to use —the classification of items is so strongly affected by the judges' own opinions as to permit the use of this procedure itself as a disguised attitude test (Sherif, Sherif, & Nebergall, 1965). The fact that, under these conditions, item placement reflects the attitudes of the judges, however, does not invalidate the use of this technique, under suitable conditions, for attitude scale construction (Hinckley, 1963; Upshaw, 1965).

LIKERT-TYPE SCALES. Another approach to the construction of attitude scales is that followed by Likert (1932). The Likert procedure does not require the classification of items by a group of judges. Items are selected solely on the basis of the responses of subjects to whom they are administered in the course of developing the test. Internal consistency is often the only criterion for item selection, although external criteria may be employed when available.

The Likert-type scale, moreover, calls for a graded response to each statement. The response is usually expressed in terms of the following five categories: strongly agree (SA), agree (A), undecided (U), disagree

(D), and strongly disagree (SD). The individual statements are either clearly favorable or clearly unfavorable. To score the scale, the alternative responses are credited 5, 4, 3, 2, or 1, respectively, from the favorable to the unfavorable end. For example, "strongly agree" with a favorable statement would receive a score of 5, as would "strongly disagree" with an unfavorable statement. The sum of the item credits represents the individual's total score, which must be interpreted in terms of empirically established norms.

An example of a modified Likert-type scale is the Minnesota Teacher Attitude Inventory. Designed to assess pupil-teacher relations, this test was developed by administering over 700 items to 100 teachers nominated by their principals as superior in pupil-teacher relations and 100 nominated as inferior. Cross validation of the resulting 150-item inventory in different groups yielded concurrent validity coefficients of .46 to .60 with a composite criterion derived from principal's estimate, pupils' ratings, and evaluation by a visiting expert. Two sample items from this inventory are shown below:[6]

> Most pupils are resourceful when left on their own.
>
> A teacher should never acknowledge his ignorance of a topic in the presence of his pupils.

For each statement, respondents mark SA, A, U, D, or SD. The scores assigned to these responses are based on criterion keying, as in the SVIB, rather than on arbitrary 1 to 5 weights. Since its publication, this test has been widely used in research. For practical application in selection and counseling, however, more information about the test is needed, especially with regard to its predictive validity and the interpretation of norms from different groups.

Several other techniques of attitude-scale construction have been employed, such as forced choice and paired comparisons; some of these techniques embody considerable theoretical sophistication and statistical refinement (see A. L. Edwards, 1957b; Fishbein, 1967; Gulliksen, 1974, pp. 254–255; Shaw & Wright, 1967; Summers, 1970). The large majority of attitude scales, however, have been developed by either the Thurstone or the Likert technique. An extensive collection of attitude scales constructed for a variety of purposes has been assembled in a book by Shaw and Wright (1967). This book provides a valuable source of items and scales for the research worker. Scale values, scoring procedures, reliability, and other technical information about each scale are also included. Few available scales, however, are ready for operational use, especially for individual assessment. Although most of the scales are moderately reliable, little information about their validity has been accumulated. Normative data are also meager. Most attitude scales must be regarded as research instruments. Information about more recently developed measures of attitudes and values, including such areas as alienation and anomie, self-esteem, and internality-externality, can be found in *Measures of Social Psychological Attitudes* (Robinson & Shaver, 1973).

---

[6] Reproduced by permission. Copyright 1951 by The Psychological Corporation, New York, N.Y. All rights reserved.

ASSESSMENT OF VALUES AND RELATED VARIABLES

Since the 1960s, there has been a resurgence of interest in the assessment of values and evaluative orientations (see, e.g., L. V. Gordon, 1975; Holtzman, 1975a). The instruments designed for this purpose vary widely in methodology, content, and specific objectives. They have much in common with measures of interests and attitudes. Some also overlap tests of other personality variables, discussed in earlier and later chapters. To illustrate the scope and variety of such instruments, four examples have been chosen.

STUDY OF VALUES. A widely used and viable early instrument in this category is the Study of Values, prepared by Allport, Vernon, and Lindzey. Originally suggested by Spranger's *Types of Men* (1928), this inventory was designed to measure the relative strength of six basic interests, motives, or evaluative attitudes, as described below:

Theoretical: Characterized by a dominant interest in the discovery of truth and by an empirical, critical, rational, "intellectual" approach.

Economic: Emphasizing useful and practical values; conforming closely to the prevailing stereotype of the "average American businessman."

Aesthetic: Placing the highest value on form and harmony; judging and enjoying each unique experience from the standpoint of its grace, symmetry, or fitness.

Social: Originally defined as love of people, this category has been more narrowly limited in later revisions of the test to cover only altruism and philanthropy.

Political: Primarily interested in personal power, influence, and renown; not necessarily limited to the field of politics.

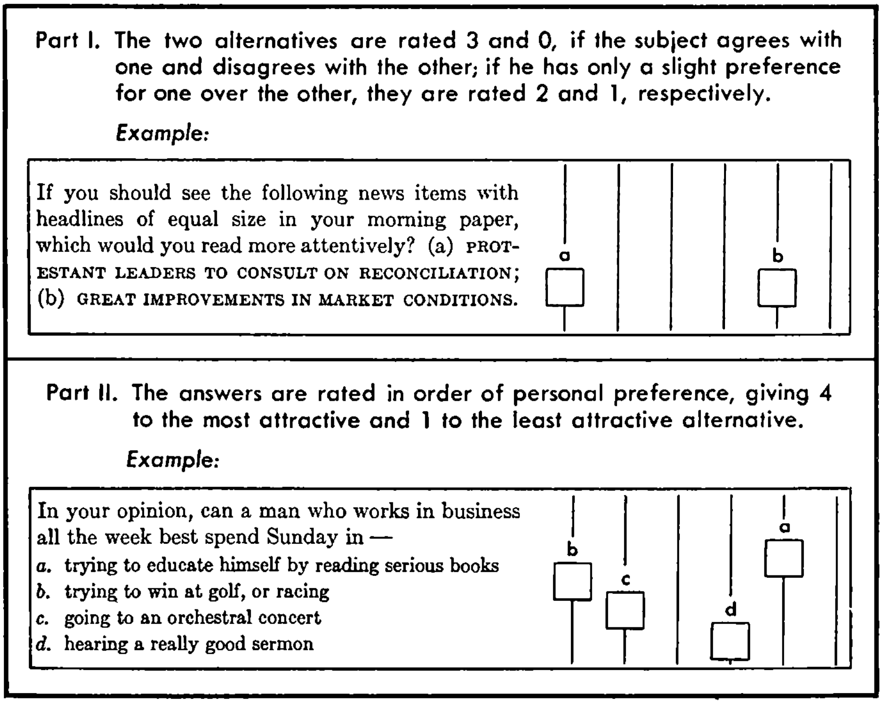
Religious: Mystical, concerned with the unity of all experience, and seeking to comprehend the cosmos as a whole.

The overlap of these types with both Holland's occupational themes and Kuder's interest areas is quite evident.

Items for the Study of Values were first formulated on the basis of the theoretical framework provided by Spranger. The criterion for the final item selection was internal consistency within each of the six areas. The items are arranged in random order in the test booklet, with no clue regarding the categories according to which they will be scored. Each item requires the preferential rating of either two or four alternatives falling in different value categories. Two sample items are reproduced in Figure 100.

Total raw scores on each of the six values are plotted in a profile. Although normative data are provided for comparative purposes, the authors clearly recognize the ipsative nature of these scores and do not recommend the use of percentiles or other normative types of scores. Mean scores in each value are given for each sex within high school and college populations, as well as for different types of colleges and for several occupational groups. The split-half reliabilities of the six scores range from .84 to .95. Retests after one or two months yielded reliabilities between .77 and .93 for the six scales.

Validity has been checked partly by the method of contrasted groups.



Part I. The two alternatives are rated 3 and 0, if the subject agrees with one and disagrees with the other; if he has only a slight preference for one over the other, they are rated 2 and 1, respectively.

Example:

If you should see the following news items with headlines of equal size in your morning paper, which would you read more attentively? (a) PROTESTANT LEADERS TO CONSULT ON RECONCILIATION; (b) GREAT IMPROVEMENTS IN MARKET CONDITIONS.

Part II. The answers are rated in order of personal preference, giving 4 to the most attractive and 1 to the least attractive alternative.

Example:

In your opinion, can a man who works in business all the week best spend Sunday in —

a. trying to educate himself by reading serious books
b. trying to win at golf, or racing
c. going to an orchestral concert
d. hearing a really good sermon

FIG. 100. Sample Items from Allport-Vernon-Lindzey Study of Values. (Reproduced by permission of Houghton Mifflin Company.)

Profiles of various educational and occupational samples exhibit significant differences in the expected directions. For example, medical students obtained their highest scores in the theoretical area, theological students in the religious area. Some relationship has been demonstrated between value profiles and academic achievement, especially when relative achievement in different fields is considered. Data are also available on the correlation of value scores with self-ratings and associates' ratings. Other overt behavioral indices of attitudes with which scores on the Study of Values have been compared include newspaper reading, descriptions of one's "ideal person," club membership, church attendance, and the like. Significant relationships in the expected directions have likewise been reported with a number of other tests, such as SVIB and Thurstone attitude scales. Finally, some studies have shown significant changes in score following specific types of experience, such as a period of study under different styles of education.

WORK VALUES INVENTORY. A more specifically focused objective guided the development of the Work Values Inventory by Super. Designed for use in academic or vocational guidance of high school and college students, as well as in personnel selection, this self-report inventory explores the sources of satisfaction the individual seeks in his work. The respondent rates each of 45 work values on a five-point scale to indicate how important it is for him (e.g., help others, can get a raise, make your own decisions).

The Work Values Inventory yields 15 scores, such as creativity, intellectual stimulation, associates, economic return, security, prestige, and altruism. Percentile norms by grade and sex are reported for a representative national sample of over 10,000 students in Grades 7 to 12. Occupational norms are being accumulated. Although there are only three items for each of the 15 values, an exploratory study with a high school sample yielded a median retest reliability of .83 for the scale scores over a two-week interval. Published in 1970, this inventory represents the culmination of some 20 years of research. Hence, the available validation data were gathered largely with earlier forms. Some suggestive findings on construct and concurrent validity are cited in the manual. Further research with the current form will provide a better basis for evaluating its effectiveness.

MORAL JUDGMENT SCALE. Although not providing any published or standardized instruments, the approach proposed by Kohlberg (1969, 1974) for assessing moral development has aroused interest among research workers. Following a Piagetian orientation, Kohlberg describes six stages of moral development, ranging from a premoral level, through a morality of conventional conformity, to the formulation of self-accepted moral principles. Kohlberg offers a cognitive model of moral development, based on the premise that individuals prefer the highest stage they can comprehend. The Moral Judgment Scale utilizes nine hypothetical dilemmas, presented one at a time. A typical example involves a man whose wife is dying of cancer but might be saved by a new drug for which an exorbitant price is demanded. Unable to raise the money, the husband steals the drug. The respondent judges the behavior described and states reasons for his judgment. In typical Piagetian "methode clinique," the interviewer elicits fuller explanations through probing questions.

Kohlberg recommends that his Moral Judgment Scale be used primarily to enhance the individual's own self-understanding and to stimulate his development in the direction he is already moving. As a research instrument, however, the scale has psychometric weaknesses that limit the generalizability of its findings (Kurtines & Greif, 1974). Both administration and scoring need to be more highly standardized to ensure comparability of results from different investigations. No data are given on short-term stability of the individual's stage of moral judgment as assessed by the scale. Nor is there evidence of internal consistency among responses to the different dilemmas. Yet these responses are combined to yield an overall index of the maturity of moral judgment reached by the individual. Like Piaget, Kohlberg argues for an invariant sequence, or ordinality, of his stages. Thus, the attainment of a higher stage should depend on the attainment of each of the preceding stages. The evidence for such invariance, however, is meager and often questionable. In summary, while the Kohlberg scale has heuristic value in stimulating discussion in a training program, it does not appear to be an effective research instrument in its present form.

INTERNAL-EXTERNAL LOCUS OF CONTROL (I-E SCALE). Another unpublished instrument, which has been extensively employed in research on a variety of problems, is the scale developed by Rotter (1966) to assess generalized expectancies for internal versus external control of reinforcement. This instrument was constructed within the context of social-learning theory (Ch. 17). In explaining its use, Rotter wrote, "The effect of reinforcement following some behavior . . . is not a simple stamping-in process but depends upon whether or not the person perceives a causal relationship between his own behavior and the reward" (1966, p. 1). Internal control refers to the individual's perception of an event as contingent upon his own behavior or his own relatively permanent characteristics. External control, on the other hand, indicates that a positive or negative reinforcement following some action of the individual is perceived as not being entirely contingent upon his own action but the result of chance, fate, or luck; or it may be perceived as under the control of powerful others and unpredictable because of the complexity of forces surrounding the individual.

Rotter refers to several earlier I-E scales prepared by other investigators and describes the 29-item scale that emerged from this research. This scale is a forced-choice self-report inventory. Two illustrative items are given in Table 33. The complete list of items, together with standard instructions for administering them, can be found in Rotter's monograph (1966). Considerable information is available regarding the I-E scale. Split-half and Kuder-Richardson reliabilities of the 29-item scale cluster around .70. Retest reliabilities after intervals of one to two months are at the same level but vary somewhat with the length of the interval, conditions of administration, and nature of the group. Correlations with social desirability scores and interest tests are low. Preliminary percentile norms on several hundred male and female students from a single university are given, as well as means and *SD*'s of some dozen other samples comprising mostly college groups.

A substantial body of data on construct validity has been accumulated. Factorial analyses indicate that a single general factor accounts for most of the variance in the responses. Other factor analyses of modified versions, however, suggest that the variable may be subdivided into several distinguishable factors, as illustrated by a belief in a difficult world, an

TABLE 33
Illustrative Items from Internal-External (I-E) Scale

(From Rotter, 1966, pp. 11, 26. Copyright 1966 by the American Psychological Association. Reprinted by permission.)

(a) In the long run, people get the respect they deserve in this world.
(b) Unfortunately, an individual's worth often passes unrecognized no matter how hard he tries.

(a) Becoming a success is a matter of hard work; luck has little or nothing to do with it.
(b) Getting a good job depends mainly on being in the right place at the right time.

Instructions state: "This is a questionnaire to find out the way in which certain important events in our society affect different people. . . . Please select the one statement of each pair (*and only one*) which you more strongly *believe* to be the case as far as you're concerned."

unjust world, an unpredictable world, and a politically unresponsive world (Collins, 1974). Comparable scales have been developed for school-age and preschool children and for minority groups (Gurin, Gurin, Lao, & Beattie, 1969; Mischel, Zeiss, & Zeiss, 1974; Nowicki & Duke, 1974; Stephens & Delys, 1973). From a psychometric viewpoint, all of these I-E inventories have been carefully constructed and evaluated.

CHAPTER 19

Projective Techniques

THE AVAILABLE supply of projective techniques is abundant and still growing. In this chapter, we shall consider the major varieties of such techniques, together with some well-known examples. Except for special points peculiar to particular techniques, no critical evaluation of individual instruments will be undertaken. Instead, a summary evaluation of projective techniques will be given in a separate section, with emphasis on common methodological problems. Projective techniques present a curious discrepancy between research and practice. When evaluated as psychometric instruments, the large majority make a poor showing. Yet their popularity in clinical use continues unabated. The nature and implications of this inconsistency will be examined in the last section.

The literature on projective techniques is vast, running to over 4,000 references on a single instrument. For a broader coverage of available projective techniques, as well as for a fuller discussion of individual instruments, the reader is referred to such sources as Holzberg (1966), Lindzey (1961), Murstein (1963), Rabin (1968), Shneidman (1965), and Zubin, Eron, and Schumer (1965). The *Mental Measurements Yearbooks* contain a separate section on projective techniques, in which multiple reviews are given for some of the more popular instruments.

NATURE OF PROJECTIVE TECHNIQUES

A major distinguishing feature of projective techniques is to be found in their assignment of a relatively *unstructured* task, i.e., a task that permits an almost unlimited variety of possible responses. In order to allow free play to the individual's fantasy, only brief, general instructions are provided. For the same reason, the test stimuli are usually vague or ambiguous. The underlying hypothesis is that the way in which the individual perceives and interprets the test material, or "structures" the situation, will reflect fundamental aspects of his psychological functioning. In other words, it is expected that the test materials will serve as a sort of screen on which the respondent "projects" his characteristic thought processes, needs, anxieties, and conflicts.

Typically, projective instruments also represent *disguised* testing procedures, insofar as the examinee is rarely aware of the type of psychological interpretation that will be made of his responses. Projective techniques are likewise characterized by a *global* approach to the appraisal of personality. Attention is focused on a composite picture of the whole personality, rather than on the measurement of separate traits. Finally, projective techniques are regarded by their exponents as especially effective in revealing *covert, latent,* or *unconscious* aspects of personality. Moreover, the more unstructured the test, it is argued, the more sensitive it is to such covert material. This follows from the assumption that the more unstructured or ambiguous the stimuli, the less likely they are to evoke defensive reactions on the part of the respondent.

Projective methods originated within a clinical setting and have remained predominantly a tool for the clinician. Some have evolved from therapeutic procedures (such as art therapy) employed with psychiatric patients. In their theoretical framework, most projective techniques reflect the influence of psychoanalytic concepts. There have also been scattered attempts to lay a foundation for projective techniques in stimulus-response theory and in perceptual theories of personality (see Lindzey, 1961, Ch. 4). It should be noted, of course, that the specific techniques need not be evaluated in the light of their particular theoretical slants or historical origins. A procedure may prove to be practically useful or empirically valid for reasons other than those initially cited to justify its introduction.

In line with their typically global approach, projective techniques have been concerned not only with emotional, motivational, and interpersonal characteristics, but also with certain intellectual aspects of behavior. Examples of the latter include general intellectual level, originality, and problem-solving styles. Certain adaptations of projective techniques have been specially designed for the measurement of attitudes (D. T. Campbell, 1950; Weschler & Bernberg, 1950), and thus supplement the instruments described in Chapter 18. It might be added that any psychological test, regardless of the purpose for which it was designed, may serve as a projective instrument. Intelligence tests, for example, are employed in this fashion by some clinicians (Fromm, Hartman, & Marschak, 1957).

INKBLOT TECHNIQUES

THE RORSCHACH. One of the most popular projective techniques is that employing the Rorschach inkblots. Developed by the Swiss psychiatrist, Hermann Rorschach (1942), this technique was first described in 1921. Although standardized series of inkblots had previously been utilized by psychologists in studies of imagination and other functions, Rorschach was the first to apply inkblots to the diagnostic investigation of the personality as a whole. In the development of this technique, Rorschach experimented with a large number of inkblots which he administered to different psychiatric groups. As a result of such clinical observations, those response characteristics that differentiated between the various psychiatric syndromes were gradually incorporated into the scoring system. The scoring procedures were further sharpened by supplementary testing of mental retardates, normals, artists, scholars, and other persons of known characteristics. Rorschach's methodology thus represented an early, informal, and relatively subjective application of criterion keying.

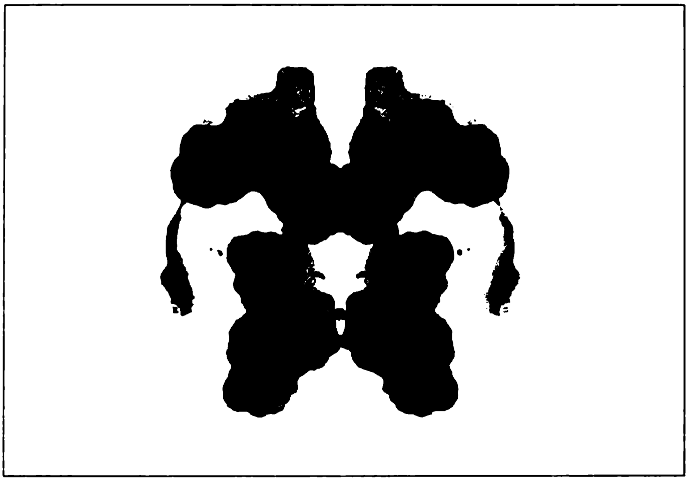


FIG. 101. An Inkblot of the Type Employed in the Rorschach Technique.

The Rorschach utilizes 10 cards, on each of which is printed a bilaterally symmetrical inkblot similar to that illustrated in Figure 101. Five of the blots are executed in shades of gray and black only; two contain additional touches of bright red; and the remaining three combine several pastel shades. As the examinee is shown each inkblot, he is asked to tell what he sees—what the blot could represent. Besides keeping a verbatim record of the responses to each card, the examiner notes time of responses, position or positions in which cards are held, spontaneous remarks, emotional expressions, and other incidental behavior of the examinee during the test session. Following the presentation of all 10 cards, the examiner questions the individual systematically regarding the parts and aspects of each blot to which the associations were given. During this inquiry, the examinee also has an opportunity to clarify and elaborate his earlier responses.

While several systems for scoring and interpreting the Rorschach have been developed, the most common scoring categories include location, determinants, and content. *Location* refers to the part of the blot with which the examinee associates each response. Does he use the whole blot, a common detail, an unusual detail, white space, or some combination of these areas? The *determinants* of the response include form, color, shading, and "movement." Although there is of course no movement in the blot itself, the respondent's perception of the blot as a representation of a moving object is scored in this category. Further differentiations are made within these categories. For example, human movement, animal movement, and abstract or inanimate movement are separately scored. Similarly, shading may be perceived as representing depth, texture, hazy forms such as clouds, or achromatic reproductions of colors as in a photograph.

The treatment of *content* varies from one scoring system to another, although certain major categories are regularly employed. Chief among these are human figures, human details (or parts of human figures), animal figures, animal details, and anatomical diagrams. Other broad scoring categories include inanimate objects, plants, maps, clouds, blood, x-rays, sexual objects, and symbols. A *popularity* score is often found on the basis of the relative frequency of different responses among people in general. For each of the 10 cards, certain responses, if given by the examinee, are scored as popular because of their common occurrence.

Further analysis of Rorschach responses is based on the relative number of responses falling into the various categories, as well as on certain ratios and interrelations among different categories. Examples of the sort of qualitative interpretations that have commonly been utilized with Rorschach responses include the association of "whole" responses with conceptual thinking, of "color" responses with emotionality, and of "human movement" responses with imagination and fantasy life. In the usual application of the Rorschach, major emphasis is placed on the final "global" description of the individual, in which the clinician integrates the results from different parts of the protocol and takes into account the interrelations of different scores and indices. In actual practice, information derived from outside sources, such as other tests, interviews, and case history records, is also utilized in preparing these global descriptions.

Some of the underlying assumptions of traditional Rorschach scoring have been called into question by a growing body of research findings. Comparative studies with standard and achromatic series of Rorschach cards, for example, have demonstrated that color itself has no effect on most of the response characteristics customarily attributed to it (Baughman, 1958). There is also evidence that verbal aptitude influences several Rorschach scores commonly interpreted as indicators of personality traits

(Lotsof, 1953; Sachman, 1952). In one study (Sachman, 1952), an analysis of the verbal complexity of individual Rorschach responses given by 100 persons revealed that "movement" responses tended to be longer and linguistically more complex than "form" responses. It was also found that verbal complexity of Rorschach responses was highly correlated with the subjects' scores on a verbal aptitude test, as well as with age and educational level.

A major complicating factor in the interpretation of Rorschach scores is the total number of responses—known as response productivity, or R (Fiske & Baughman, 1953). Because of large individual differences in R, the practice of considering the absolute number of responses in various categories is obviously misleading. If two individuals or groups differ in R, they are also likely to differ in the same direction in the number of responses falling in specific categories. Thus, the differences found in certain categories may be only an artifact resulting from the variation in total number of responses. Nor is the use of percentages a completely satisfactory solution. For example, a protocol containing many responses is likely to include a smaller *proportion* of "whole" responses, since the number of "whole" responses that can reasonably be perceived in the blots is quite limited, while associations to isolated details of the blots may continue indefinitely.

To these intrinsic characteristics of the Rorschach scores may be added the empirical fact that response productivity appears to be closely related to age, intellectual level, and amount of education. Even more disturbing is the finding that R varies significantly from one examiner to another (Baughman, 1951; Guilford & Lacey, 1947). All of these results suggest that R, which may itself be a major determinant of many of the common Rorschach scores, is influenced by factors quite extraneous to the basic personality variables allegedly measured by the Rorschach. Findings such as these strike at the very foundation on which the entire elaborate superstructure of Rorschach scoring is supported.

Within the mass of negative and inconclusive research findings about the Rorschach, a few encouraging trends can be noted. A handbook by Goldfried, Stricker, and Weiner (1971) presents a systematic survey of available data on norms, reliability, and validity for a number of specific applications of the Rorschach technique in both research and clinical practice. Examples of these applications include proposed sets of scoring signs for developmental level, hostility, anxiety, and brain dysfunction. Each application is accompanied by a thorough and well-balanced evaluation of its strengths and weaknesses. The authors conclude that, while several of the Rorschach indices surveyed have demonstrated enough validity to justify their use in research, their suitability for clinical use has not been sufficiently established. They offer methodological suggestions for the kinds of research needed to provide adequate evaluation of each technique. They also conclude that the Rorschach appears most promising when used as a direct measure of cognitive style and perceptual organization, or as a structured interview in the hands of a skilled clinician. Results are least encouraging when the Rorschach is used to elicit fantasy production for psychoanalytic and other symbolic interpretations.

Although the Rorschach is considered to be applicable from the preschool to the adult level, its normative data were originally derived in large part from adult groups. This limitation also characterizes the general fund of clinical experience accumulated through the use of the Rorschach and employed in the qualitative interpretation of protocols. In the effort to extend the empirical framework for Rorschach interpretation to other age groups, Ames and her co-workers at the Gesell Institute of Child Development at Yale collected and published Rorschach norms on children between the ages of 2 and 10 years, on adolescents between the ages of 10 and 16, and on older persons from the age of 70 up (Ames, Learned, Metraux, & Walker, 1954; Ames, Metraux, Rodell, & Walker, 1974; Ames, Metraux, & Walker, 1971). A comprehensive and integrated summary of published normative data for children and adolescents was prepared by Levitt and Truumaa (1972). Combining data from 15 studies, the authors provide means of various quantitative Rorschach indices by intellectual level and by age up to 16 years.

Mention may also be made of recent efforts to apply computer techniques to the interpretation of Rorschach protocols. A computerized scoring and interpretive system, known as Piotrowski's Automated Rorschach (PAR), became available for operational use in 1974. Based on the research and clinical observations of Piotrowski (1964), this system yields a personality description consisting of some 50 to 80 statements per protocol. Employing 340 Rorschach response indicators, the computer selects these statements from a pool of 980 messages. The statements pertain to such characteristics as interpersonal skills, intellectual efficiency, initiative and autonomy, problem-solving skills, and personality problems and their possible causes. The reports are differently organized and adapted for use in different contexts, such as industrial organizations, correctional institutions, and clinics.

THE HOLTZMAN INKBLOT TECHNIQUE. Unlike most projective devices, the Holtzman Inkblot Technique (HIT) represents a genuine attempt to meet the technical standards of psychometric instruments (Holtzman, 1968, 1975b; Holtzman, Thorpe, Swartz, & Herron, 1961). Modeled after the Rorschach, the Holtzman test was so designed as to eliminate the principal technical deficiencies of the earlier instrument. The changes in stimulus materials and procedure are sufficiently extensive, however, to require that the Holtzman be regarded as a new test and evaluated without reference to either the theoretical or empirical characteristics of the Rorschach. The Holtzman technique provides two parallel series of 45 cards each. Only one response per card is obtained. Both achromatic and colored cards are included; a few inkblots are markedly asymmetric. The blots were selected from a large preliminary pool on the basis of three empirical criteria: (1) discrimination between a group of college student volunteers and a group of hospitalized psychotics; (2) "pulling power" of the blot in eliciting responses scorable on such traditional Rorschach variables as location, color, shading, and movement; and (3) scorer reliability.

Administration and scoring of the HIT are well standardized and clearly described. Scores are obtained in 22 response variables, including most of the Rorschach form variables and such additional variables as pathognomic verbalization, anxiety, and hostility. For each variable, percentile scores are reported for several normal samples ranging from 5-year-olds to adults, and for a number of deviant groups, such as schizophrenics, mental retardates, emotionally disturbed children, juvenile delinquents, and alcoholics (Hill, 1972). Although these norms were derived from nearly 2,000 persons, the number of cases in most of the groups is small.

Scorer reliability appears to be highly satisfactory. Both split-half and alternate-form reliability have been extensively investigated, with generally encouraging results. A group form of the test, using slides, yields scores on most variables that are closely similar to those obtained through individual administration (Holtzman, Moseley, Reinehr, & Abbott, 1963). A short, 30-item version of the group form also gives similar results (Herron, 1963). Computer scoring is available for the group form (Gorham, 1967); and work is in progress on the development of a computer-based interpretive system (Holtzman, 1975b).

It is apparent that the HIT has many psychometric advantages over the Rorschach. The availability of parallel forms permits not only the measurement of retest reliability but also follow-up studies with alternate forms. The restriction of responses to one per card holds response productivity (R) constant for each examinee, thus avoiding many of the pitfalls of Rorschach scoring. It should be noted, however, that response length (number of words) is still uncontrolled and, as in the Rorschach, has proved to be significantly related to several HIT scores, notably the frequency of movement responses (Megargee, 1966a).

Considerable validity data on the HIT have been accumulated (Gamble 1972; Holtzman, 1968, 1975b). The validation research has followed a variety of approaches, including the study of developmental trends, cross-cultural comparisons, correlation with other tests and with behavioral indicators of personality characteristics, and contrasted group comparisons with both normal and pathological subjects. Available results are promising; but more data are needed to establish the diagnostic significance of the various scores and the construct validity of the personality variables assessed by this technique. A handbook prepared by Hill (1972) is directed particularly to the clinical use of the HIT. Integrating research findings on the HIT with the accumulated Rorschach literature and clinical experience, Hill provides a guide for interpreting HIT protocols on two levels. The first level is concerned with the 22 quantitatively scored variables and their interrelations. The second represents a deeper interpretive level in terms of broader and more highly integrated personality variables. Three major descriptive categories are employed: cognitive functioning (intellectual level and thought disturbances), affective functioning (e.g., levels of maturity, forms of anxiety, defense mechanisms), and self-concept (including several forms of sexual identity). The handbook also provides five detailed case studies.

THE THEMATIC APPERCEPTION TEST AND RELATED INSTRUMENTS

THEMATIC APPERCEPTION TEST. In contrast to inkblot techniques, the Thematic Apperception Test (TAT) presents more highly structured stimuli and requires more complex and meaningfully organized verbal responses. Interpretation of responses by the examiner is usually based on content analysis of a rather qualitative nature. First developed by Murray and his staff at the Harvard Psychological Clinic (Murray et al., 1938), the TAT has not only been widely used in clinical practice and research, but it has also served as a model for the development of many other instruments (J. W. Atkinson, 1958; Harrison, 1965; Henry, 1956; Varble, 1971).

The TAT materials consist of 19 cards containing vague pictures in black and white and one blank card. The examinee is asked to make up a story to fit each picture, telling what led up to the event shown in the picture, describing what is happening at the moment and what the characters are feeling and thinking, and giving the outcome. In the case of the blank card, he is instructed to imagine some picture on the card, describe it, and then tell a story about it. The original procedure outlined by Murray in the test manual requires two one-hour sessions, 10 cards being employed during each session. The cards reserved for the second session were deliberately chosen to be more unusual, dramatic, and bizarre, and the accompanying instructions urge the individual to give free play to his imagination. Four overlapping sets of 20 cards are available—for boys, girls, men over 14, and women over 14. Most clinicians use abridged sets of specially selected cards, seldom giving more than 10 cards to a single respondent. A card from the second set of the adult women's series is shown in Figure 102.

In interpreting TAT stories, the examiner first determines who is the "hero," the character of either sex with whom the subject has presumably identified himself. The content of the stories is then analyzed principally in reference to Murray's list of "needs" and "press." Several of the proposed needs were described in Chapter 17, in connection with the Edwards Personal Preference Schedule and the Personality Research Form. Examples include achievement, affiliation, and aggression. Press refers to environmental forces that may facilitate or interfere with the satisfaction of needs. Being attacked or criticized by another person, receiving affection, being comforted, and exposure to physical danger as in a shipwreck are illustrations of press. In assessing the importance or strength of a particular need or press for the individual, special attention is given to the intensity, duration, and frequency of its occurrence in different stories, as well as to the uniqueness of its association with a given picture. The assumption is made that unusual material, which departs from the common responses to each picture, is more likely to have significance for the individual.

A fair amount of normative information has been published regarding the most frequent response characteristics for each card, including the

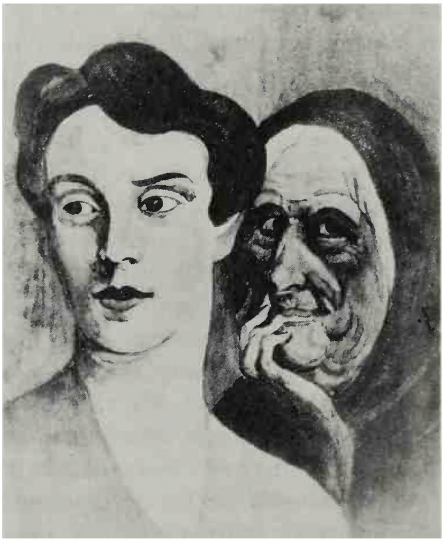


Fig. 102. One of the Pictures Used in the Thematic Apperception Test.
(Reproduced by permission of Harvard University Press.)

way each card is perceived, the themes developed, the roles ascribed to the characters, emotional tones expressed, speed of responses, length of stories, and the like (see J. W. Atkinson, 1958; Henry, 1956). Although these normative data provide a general framework for interpreting individual responses, most clinicians rely heavily on "subjective norms" built up through their own experience with the test. A number of quantitative scoring schemes and rating scales have been developed that yield good scorer reliability. Since their application is rather time consuming, however, such scoring procedures are seldom used in clinical practice. Although typically given as an individual oral test in the clinical situation, the TAT may also be administered in writing and as a group test. There is some evidence suggesting that under the latter conditions productivity of meaningful material may be facilitated.

The TAT has been used extensively in personality research. Several investigations have been concerned with the assumptions that underlie TAT interpretations, such as self-identification with the hero and personal significance of uncommon responses (Lindzey, 1959). Although they cannot establish criterion-related validity of the TAT for specific uses, such studies contribute to the construct validation of TAT interpretations. A basic assumption that TAT shares with other projective techniques is that present motivational and emotional condition of the subject affects his responses to an unstructured test situation. A considerable body of experimental data is available to show that such conditions as hunger, sleep deprivation, social frustration, and the experience of failure in a preceding test situation significantly affect TAT responses (J. W. Atkinson, 1958). While supporting the projective hypothesis, the sensitivity of the TAT to such temporary conditions may complicate the detection of more enduring personality traits. Interpretation of the constructs assessed by the TAT must also take into account the finding that measures of the same needs through such instruments as the TAT, EPPS, and Adjective Check List have shown little correspondence of results (Megargee & Parker, 1968). It is possible that these different techniques assess different aspects of the common needs. In any event, such discrepancies limit the generalizability of scores. The question of internal consistency of TAT responses must also be considered in this connection (Entwisle, 1972). And some control should be applied for length of stories, or productivity—a problem that the TAT shares with the Rorschach (Veroff, Atkinson, Feld, & Gurin, 1974).

ADAPTATIONS OF THE TAT AND RELATED TESTS. Many adaptations of the TAT have been developed for special purposes. These exhibit varying degrees of resemblance to the original. Where to draw the line between modified versions of the TAT and new tests based on the same general approach as the TAT is arbitrary. Several versions of the TAT have been prepared for use in attitude surveys; examples include sets for assessing attitudes toward labor problems, minority groups, school, and authority (D. T. Campbell, 1950; Harrison, 1965). Other adaptations have been developed for use in vocational counseling, executive appraisal, and a wide variety of research projects. Forms have been constructed for special populations, including preschool children, elementary schoolchildren, adolescents, crippled children, and various national and ethnic groups (Harrison, 1965).

Some TAT adaptations have focused on the intensive measurement of a single need or drive, such as sex or aggression. Of special interest is the extensive research on the achievement need (n-Ach) conducted by McClelland, J. W. Atkinson, and their associates (J. W. Atkinson, 1958; Atkinson, Feather, et al., 1966; McClelland, Atkinson, Clark, & Lowell 1953). To measure n-Ach, four pictures were employed, two of which were taken from the TAT. The cards portray men working at a machine, a boy at a desk with a book, a father and son picture, and a boy who is apparently daydreaming. Detailed scoring schemes have been developed for scoring the resulting stories with regard to expressions of n-Ach. This technique has been utilized in an extensive program of research on achievement motivation. The problems investigated range from basic motivation theory (Atkinson, Feather, et al. 1966) to the social origins and consequences of n-Ach and its role in the rise and fall of societies (McClelland, 1961). Still another special set of cards (six in a male form and six in a female form) was developed for use in a national survey of three needs: achievement, affiliation, and power (Veroff et al., 1974). For this purpose, the instructions to respondents were more highly standardized; and responses were given orally and recorded by the interviewers.

Although the original TAT is said to be applicable to children as young as 4 years of age, the Children's Apperception Test (CAT) was specially designed for use between the ages of 3 and 10 years (Bellak, 1975). The CAT cards substitute animals for people on the assumption that young children project more readily to pictures of animals than to pictures of humans. The various animals in the pictures are portrayed in typically human situations, in the characteristic anthropomorphic fashion of comic strips and children's books. The pictures are designed to evoke fantasies relating to problems of feeding and other oral activity, sibling rivalry, parent-child relations, aggression, toilet-training, and other childhood experiences. Contrary to the authors' assumption, several studies with children from the first grade up found either no difference or, more often, greater productivity of clinically significant material with human than with animal pictures (Murstein, 1963). In response to these research findings, the authors of the CAT prepared a human modification of the test

(CAT-H) for use with older children, especially those with a mental age beyond 10 years (Bellak & Hurvich, 1966). The authors maintain that either the human or the animal form may be more effective depending on the age and personality characteristics of the child. Similar thematic apperception tests have been developed for the aged, including the Gerontological Apperception Test (Wolk & Wolk, 1975) and the Senior Apperception Technique (Bellak, 1975; Bellak & Bellak, 1973). Both employ sets of cards featuring one or more elderly persons and illustrating problems of concern to the aged, such as loneliness, family difficulties, dependence, and helplessness.

OTHER PROJECTIVE TECHNIQUES

VERBAL TECHNIQUES. Although all projective instruments discussed thus far require verbal responses, certain projective techniques are wholly verbal, utilizing only words in both stimulus materials and responses. Some of these verbal techniques can be administered in either oral or written form; but all are suitable for written group administration. When so administered, of course, they presuppose a minimum reading level and thorough familiarity with the language in which the test was developed. These requirements thus preclude the use of such techniques with young children, illiterates, or the foreign-speaking.

A technique that antedated the flood of projective tests by more than half a century is the *word association test.* Originally known as the "free association test," this technique was first systematically described by Galton (1879). Wundt and J. McK. Cattell subsequently introduced it into the psychological laboratory, where it was adapted to many uses. The procedure involves simply the presentation of a series of disconnected words, to each of which the subject is told to respond by giving the first word that comes to his mind. The early experimental psychologists, as well as the first mental testers, saw in such association tests a tool for the exploration of thinking processes.

The clinical application of word association methods was stimulated largely by the psychoanalytic movement, although other psychiatrists, such as Kraepelin, had previously investigated such techniques. Among the psychoanalysts, Jung's contribution to the systematic development of the word association test is most conspicuous. Jung (1910) selected stimulus words to represent common "emotional complexes," and analyzed the responses with reference to reaction time, content, and physical expressions of emotional tension. Over thirty years later, a word association technique was developed at the Menninger Clinic by Rapaport and his associates (1946, Ch. 2). In its general orientation, this adaptation reveals its kinship to the earlier Jung test. The 60-word list contains a preponderance of words selected for their psychoanalytic significance, many of them being associated with psychosexual conflicts. According to its authors, the test had a dual aim: to aid in detecting impairment of thought processes and to suggest areas of significant internal conflicts. Results are analyzed with reference to such characteristics as proportion of common or popular responses, reaction time, associative disturbances, and impaired reproduction on retest.

Mention may also be made of the use of the word association technique as a "lie detector." This application was likewise initiated by Jung and was subsequently subjected to extensive research—both in the laboratory and in practical situations (Burtt, 1931; Lindsley, 1955). The rationale offered to justify the employment of word association in the detection of lying or guilt is similar to that which underlies its utilization in uncovering areas of emotional conflict. Content analysis, reaction time, and response disturbances have all been explored as indices of lying or guilt. Frequently, physiological measures of emotional excitement are obtained concurrently with the verbal responses. The word lists chosen for lie detection purposes are usually custom-made to cover distinctive features of the particular crime or other situation under investigation. Whether word association has any practical value as a lie detector is still a moot point. Its effectiveness certainly varies widely with the specific circumstances under which it is used.

A different approach to the word association test is illustrated by the early work of Kent and Rosanoff (1910). Designed principally as a psychiatric screening instrument, the Kent-Rosanoff Free Association Test utilized completely objective scoring and statistical norms. The stimulus words consisted of 100 common, neutral words, chosen because they tend to evoke the same associations from people in general. For example, to the word *table*, most people respond "chair"; to *dark*, they say "light." A set of frequency tables was prepared—one for each stimulus word—showing the number of times each response was given in a standardization sample of 1,000 normal adults.

In scoring the Kent-Rosanoff test, the median frequency value of the responses the examinee gives to the 100 stimulus words is employed as an "index of commonality." Any responses not found in the normative tables are designed as "individual." Comparisons of psychotics with normals suggested that psychotics give more individual responses and obtain a lower index of commonality than normals. The diagnostic use of the test declined, however, with the gradual realization that response frequency also varies widely with age, socioeconomic and educational level, regional and cultural background, creativity, and other factors. Hence, proper interpretation of scores requires norms on many subgroups, as well as supplementary information about the examinee. The Kent-Rosanoff test has nevertheless retained its position as a standard laboratory technique. Additional norms have been gathered in several countries, and the technique has been extensively employed in research on verbal behavior and personality (Jenkins & Russell, 1960; Palermo & Jenkins, 1963; Postman & Keppel, 1970; Van der Made-Van Bekkum, 1971).

Another verbal projective technique, *sentence completion*, has been widely employed in both research and clinical practice. Generally, the opening words, or sentence stems, permit an almost unlimited variety of possible completions. Examples might be: My ambition . . . ; Women . . . ; What worries me . . . ; My mother. . . . The sentence stems are frequently formulated so as to elicit responses relevant to the personality domain under investigation. This flexibility of the sentence-completion technique represents one of its advantages for clinical and research purposes. Nevertheless, some standardized forms have been published for more general application.

An example is the Rotter Incomplete Sentences Blank, consisting of 40 sentence stems. The directions to the examinee read: "Complete these sentences to express *your real feelings*. Try to do every one. Be sure to make a complete sentence." Each completion is rated on a 7-point scale according to the degree of adjustment or maladjustment indicated. Illustrative completions corresponding to each rating are given in the manual. With the aid of these specimen responses, fairly objective scoring is possible. The sum of the individual ratings provides a total adjustment score that can be used for screening purposes. The response content can also be examined clinically for more specific diagnostic clues.

A more specialized application of the sentence completion technique is provided by the Washington University Sentence Completion Test (WUSCT), based on Loevinger's broadly defined construct of ego development. Closely reflecting the author's theoretical framework, the WUSCT classifies the responses with reference to a seven-stage scale of ego development, or I-level, as follows: presocial and symbiotic, impulsive, self-protective, conformist, conscientious, autonomous, and integrated. Each response is assigned to a single I-level, and a composite score on the entire test is computed from these values. Although most of the research has been done with adult women, forms are also available for use with men and with younger persons of either sex. The two-volume report on the development of the WUSCT contains detailed instructions for scoring and interpretation, as well as an effective set of self-training exercises for scorers (Loevinger, Wessler, & Redmore, 1970).

PICTORIAL TECHNIQUES. Projective techniques employ pictorial material in a variety of ways. The use of pictures to stimulate the free play of fantasy and to evoke elaborate verbal responses, as in the TAT, has already been consisdered. In contrast, the instruments illustrated in this section are more circumscribed in coverage and call for simpler responses. An example is the Rosenzweig Picture-Frustration Study (P-F Study), available in a form for children (4 to 13 years) and a form for adults (14 years and over). Derived from the author's theory of frustration and aggression, this test presents a series of cartoons in which one person frustrates another. Two of these cartoons, from the children's form, are shown in Figure 103. In the blank space provided, the respondent writes what the frustrated person would reply. Responses are classified with reference to type and direction of aggression. Type of aggression includes: obstacle-dominance, emphasizing the frustrating object; ego-defense, focusing attention on the protection of the frustrated person; and need-persistence, concentrating on the constructive solution of the frustrating problem. Direction of aggression is scored as: extrapunitive, or turned outward on the environment; intropunitive, or turned inward on the subject; and impunitive, or turned off in an attempt to gloss over or evade the situation. In scoring the test, the percentage of responses falling into each of these categories is compared with the corresponding normative percentages. A group conformity rating (GCR), showing the individual's tendency to give responses that agree with the modal responses of the standardization sample, may also be obtained.

Being more limited in coverage, more highly structured, and relatively objective in its scoring procedures, the P-F Study lends itself better to statistical analysis than do most other projective techniques. Systematic

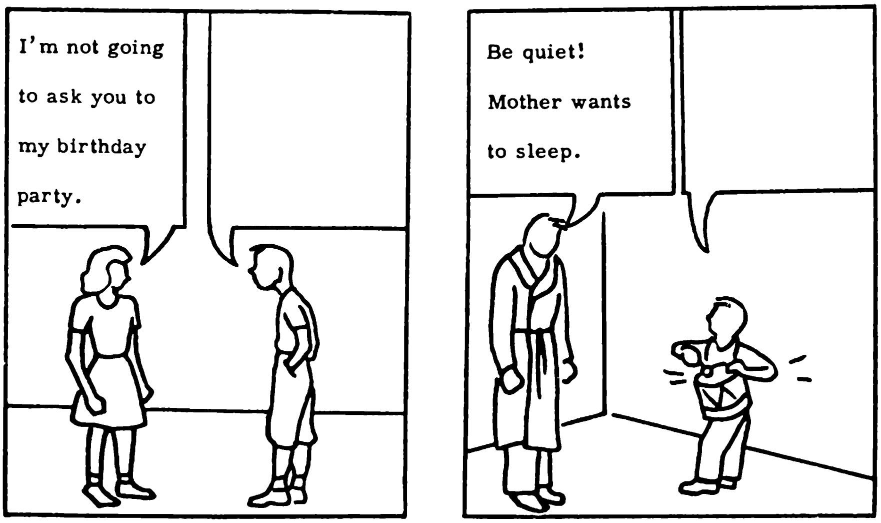

I'm not going to ask you to my birthday party.
Be quiet! Mother wants to sleep.

FIG. 103. Typical Items from the Rosenzweig Picture-Frustration Study, Children's Form.

(Reproduced by permission of Saul Rosenzweig.)

efforts have been made to gather norms and to check its reliability and validity. Like a number of other projective instruments, the P-F Study has been adapted by other investigators for research on a variety of problems. The general approach of the P-F Study has been followed, for example, in tests designed to assess attitudes toward minority groups (J. F. Brown, 1947), opinions on the prevention of war (Fromme, 1941), consumers' responses to products (G. H. Smith, 1954, pp. 120–121), and reactions of amputees toward their disability (Peizer, 1958).

Still another approach is illustrated by Animal Crackers, a recently developed test of achievement motivation in early childhood, with special reference to the school setting (Adkins & Ballif, 1975). Described by its authors as an "objective-projective" test, Animal Crackers presents pairs of identical animals in each item. The child is told that he has his "own" animals which like what he likes and do what he does. The examiner then points in turn to each animal in a pair, describing its behavior, and asks the child to identify his own animal. Two items, taken from different parts of the test, are shown in Figure 104. No verbal response is required. The child either points (in the individual administration) or fills in the appropriate circle (in the group administration).

Animal Crackers consists of 60 items, 12 chosen to measure each of the following five components of achievement-oriented school behavior (Adkins & Ballif, 1975, p. 6):

School enjoyment: the child expects pleasurable experiences from achieving in school.

Self-confidence: the child sees himself as someone who can achieve in school.

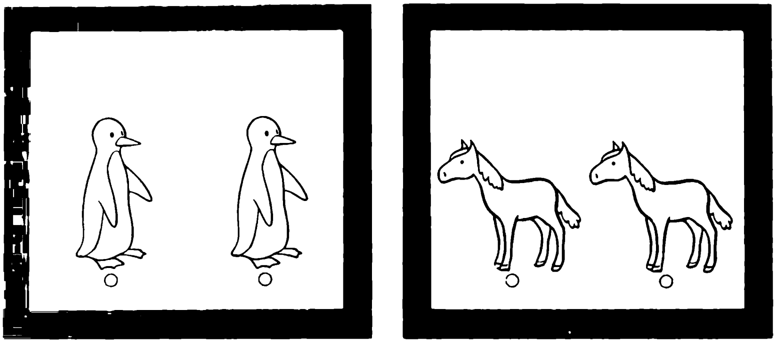
Purposiveness: the child can set up purposes and goals and can let them direct his behavior.

Instrumental activity: the child knows how to accomplish these goals by engaging in appropriate instrumental steps.

Self-evaluation: the child can evaluate his own performance in school.

Although the test uses a two-alternative forced-choice technique, the component scores are not ipsative, since the paired responses in each item refer to the same trait. The test was developed through several years of research with children from various ethnic and socioeconomic groups. Items were chosen on the basis of response frequencies, item-test correlations, and factor analysis. The design of the test incorporates controls for several response sets.

Percentile norms for kindergarten and first grade are presented on a representative national sample of over 5,000 children tested in 32 states. The available norms show a pronounced negative skewness in the score



(left) This penguin wants to learn something new today.

(right) This penguin wants to do what it learned yesterday.

(right) This pony watches the teacher show how to do something.

(left) This pony looks around when the teacher shows how to do something.

The two response options in each item are given in the order in which the examiner reads them, while pointing to the designated animal. Reversals in order and position of keyed options are introduced to control response sets.

FIG. 104. Two Items from Animal Crackers: A Test of Motivation to Achieve. (Reproduced by permission of CTB/McGraw-Hill, Inc.)

distribution, especially for the first-grade samples, suggesting inadequate test ceiling. At these levels, the test appears to be most useful in identifying children with low motivation. Among preschool children, for whom the test had originally been designed, finer discriminations can be made along the entire range. Although norms are provided for both total scores and component scores, users are cautioned about the lower reliability of component scores. Kuder-Richardson reliabilities are in the .90's for the total test, but range from .69 to .92 for the component scores. The manual notes that these values are somewhat inflated because of the ceiling effect. Data on validity are meager, except for the contribution of factor-analytic results to the construct validation of the component scores. The test appears to have promise as a research tool.

EXPRESSIVE TECHNIQUES. A large and amorphous category of projective techniques comprises many forms of relatively free self-expression. It is characteristic of all these techniques that they have been employed as therapeutic as well as diagnostic procedures. Through the opportunities for self-expression that these activities afford, it is believed that the individual not only reveals his emotional difficulties but also relieves them. The techniques most frequently employed in this category are drawing and dramatic use of toys.

Although almost every art medium, technique, and type of subject matter has been investigated in the search for significant diagnostic clues, special attention has centered on *drawings of the human figure.* A well-known example is provided by the Machover Draw-a-Person Test (D-A-P). In this test, the examinee is provided with paper and pencil, and told simply to "draw a person." Upon completion of the first drawing, he is asked to draw a person of the opposite sex from that of the first figure. While the individual draws, the examiner notes his comments, the sequence in which different parts are drawn, and other procedural details. The drawing may be followed by an inquiry, in which the examinee is asked to make up a story about each person drawn, "as if he were a character in a play or novel." A series of questions is also employed during the inquiry to elicit specific information about age, schooling, occupation, family, and other facts associated with the characters portrayed.

Scoring of the Draw-a-Person Test is essentially qualitative, involving the preparation of a composite personality description from an analysis of many features of the drawings. Among the factors considered in this connection are the absolute and relative size of the male and female figures, their position on the page, quality of lines, sequence of parts drawn, stance, front or profile view, position of arms, depiction of clothing, and background and grounding effects. Special interpretations are given for the omission of different bodily parts, disproportions, shading, amount and distribution of details, erasures, symmetry, and other stylistic features. There is also detailed discussion of the significance of each major body part, such as head, individual facial features, hair, neck, shoulders, breast, trunk, hips, and extremities.

The interpretive guide for the Draw-a-Person Test abounds in sweeping generalizations, such as "Disproportionately large heads will often be given by individuals suffering from organic brain disease," or "The sex given the proportionately larger head is the sex that is accorded more intellectual and social authority." But no evidence is provided in support of these statements. Reference is made to a file of "thousands of drawings" examined in clinical contexts, and a few selected cases are cited for illustrative purposes. No systematic presentation of data, however, accompanies the original published report of the test (Machover, 1949). Validation studies by other investigators have yielded conflicting results, the better controlled studies lending no support to the diagnostic interpretations proposed by Machover (Roback, 1968; Swensen, 1968; Watson, Felling, & Maceachern, 1967). There is some evidence suggesting that global ratings may be more reliable and valid than individual signs. However, the overall artistic quality of the drawings, reflecting largely individual differences in drawing ability, accounts for much of the variance of such global judgments.

Toy tests, involving such objects as puppets, dolls, and miniatures, have been widely utilized in projective testing. Originating in play therapy with children, these materials have subsequently been adapted for the diagnostic testing of both adults and children. The objects are usually selected because of their expected associative value. Among the articles most frequently employed for these purposes, for example, are dolls representing adults and children of both sexes, furniture, bathroom and kitchen fixtures, and other household furnishings. Play with such articles is expected to reveal the child's attitudes toward his own family, as well as sibling rivalries, fears, aggressions, conflicts, and the like. The examiner notes what items the child chooses and what he does with them, as well as his verbalizations, emotional expressions, and other overt behavior.

With children, these techniques often take the form of free play with the collection of toys that the examiner simply makes available. With adults, the materials are presented with general instructions to carry out some task of a highly unstructured nature. These instructions may, of course, also be employed with children. Frequently the task has certain dramatic features, as in the arrangement of figures on a miniature stage set. Several investigators have used play techniques in studying prejudice and other intergroup attitudes (D. T. Campbell, 1950).

EVALUATION OF PROJECTIVE TECHNIQUES

It is evident that projective techniques differ widely among themselves. Some appear more promising than others because of more favorable empirical findings, sounder theoretical orientation, or both. Regarding some techniques, such as the Rorschach, voluminous data have been gathered, although their interpretation is often uncertain. About others little is known, either because of their recent origin, or because objective verification is hindered by the intrinsic nature of the instruments or by the attitudes of their exponents.

To evaluate each instrument individually and to attempt to summarize the extensive pertinent literature would require a separate volume. Within this chapter, critical comments have been interjected only in the cases of instruments that presented unique features—whether of a favorable or unfavorable nature. There are certain points, however, that apply to a greater or lesser extent to the bulk of projective techniques. These points can be conveniently considered in summary form.

RAPPORT AND APPLICABILITY. Most projective techniques represent an effective means for "breaking the ice" during the initial contacts between examiner and examinee. The task is usually intrinsically interesting and often entertaining. It tends to divert the individual's attention away from himself and thus reduces embarrassment and defensiveness. And it offers little or no threat to his prestige, since any response he gives is "right."

Certain projective techniques may be especially useful with young children, illiterates, and persons with language handicaps or speech defects. Nonverbal media would be readily applicable to all these groups. And oral responses to pictorial and other nonlanguage stimuli could be secured from the first two. With all these verbally limited groups, projective techniques may help the examinee to communicate with the examiner. These techniques may also aid the individual in clarifying for himself some of his own behavior that he had not previously verbalized.

FAKING. In general, projective instruments are less susceptible to faking than are self-report inventories. The purpose of projective techniques is usually disguised. Even if an individual has some psychological sophistication and is familiar with the general nature of a particular instrument, such as the Rorschach or TAT, it is still unlikely that he can predict the intricate ways in which his responses will be scored and interpreted. Moreover, the examinee soon becomes absorbed in the task and hence is less likely to resort to the customary disguises and restraints of interpersonal communication.

On the other hand, it cannot be assumed that projective tests are completely immune to faking. Several experiments with the Rorschach, TAT, and other projective instruments have shown that significant differences do occur when subjects are instructed to alter their responses so as to create favorable or unfavorable impressions, or when they are given statements suggesting that certain types of responses are more desirable (Masling, 1960). In a particularly well-controlled study, Davids and Pildner (1958) administered a battery of self-report and projective tests to two groups of college students, one of which took the tests as genuine job applicants, the other as participants in a research project. Under these conditions, the job applicants obtained significantly better-adjusted scores than the research subjects on the self-report but *not* on the projective tests. Certain types of projective test items, however, were found to be susceptible to faking. For example, sentence completion stems expressed in the first person yielded significantly more favorable responses than those expressed in the third person.

EXAMINER AND SITUATIONAL VARIABLES. It is obvious that most projective techniques are inadequately standardized with respect to both administration and scoring. Yet there is evidence that even subtle differences in the phrasing of verbal instructions and in examiner-examinee relationships can appreciably alter performance on these tests (Baughman, 1951; Hamilton & Robertson, 1966; Herron, 1964; Klinger, 1966; Masling, 1960; S. B. Sarason, 1954). Even when employing identical instructions, some examiners may be more encouraging or reassuring, others more threatening, owing to their general manner and appearance. Such differences may affect response productivity, defensiveness, stereotypy, imaginativeness, and other basic performance characteristics. In the light of these findings, problems of administration and testing conditions assume even greater importance than in other psychological tests.

Equally serious is the lack of objectivity in scoring. Even when objective scoring systems have been developed, the final steps in the evaluation and integration of the raw data usually depend on the skill and clinical experience of the examiner. Such a situation has several implications. In the first place, it reduces the number of examiners who are properly qualified to employ the technique and thus limits the range of its effective application. It also means that the results obtained by different examiners may not be comparable, a fact that complicates research with the instrument. But perhaps the most disturbing implication is that the interpretation of scores is often as projective for the examiner as the test stimuli are for the examinee. In other words, the final interpretation of projective test responses may reveal more about the theoretical orientation, favorite hypotheses, and personality idiosyncrasies of the examiner than it does about the examinee's personality dynamics.

NORMS. Another conspicuous deficiency common to many projective instruments pertains to normative data. Such data may be completely lacking, grossly inadequate, or based on vaguely described populations. In the absence of adequate objective norms, the clinician falls back on his "general clinical experience" to interpret projective test performance. But such a frame of reference is subject to all the distortions of memory that are themselves reflections of theoretical bias, preconceptions, and other idiosyncrasies of the clinician. Moreover, any one clinician's contacts may have been limited largely to persons who are atypical in education, socioeconomic level, sex ratio, age distribution, or other relevant characteristics. In at least one respect, the clinician's experience is almost certain to produce a misleading picture, since he deals predominantly with maladjusted or pathological cases. He thus lacks sufficient firsthand familiarity with the characteristic reactions of normal people. The Rorschach norms gathered by Ames and her associates on children, adolescents, and persons over 70 represent one effort to correct some of the more obvious lacks in this regard.

Interpretation of projective test performance often involves subgroup norms, of either a subjective or an objective nature. Thus, the clinician may have a general subjective picture of what constitutes a "typical" schizophrenic or psychoneurotic performance on a particular test. Or the published data may provide qualitative or quantitative norms that delineate the characteristic performance of different diagnostic groups. In either case, the subgroup norms may lead to faulty interpretations unless the subgroups were equated in other respects. For example, if the schizophrenics and normals on whom the norms were derived differed also in educational level, the observed disparities between schizophrenic and normal performance may have resulted from educational inequality rather than from schizophrenia. Similar systematic or constant errors may operate in the comparison of various psychiatric syndromes. For example, schizophrenics as a group tend to be younger than manic-depressives; anxiety neurotics are likely to come from higher educational and socioeconomic levels than hysterics.

RELIABILITY. In view of the relatively unstandardized scoring procedures and the inadequacies of normative data, *scorer reliability* becomes an important consideration in projective testing. For projective techniques, a proper measure of scorer reliability should include not only the more objective preliminary scoring, but also the final integrative and interpretive stages. It is not enough, for example, to demonstrate that examiners who have mastered the same system of Rorschach scoring agree closely in their tallying of such characteristics as whole, unusual detail, or color responses. On a projective test like the Rorschach, these raw quantitative measures cannot be interpreted directly from a table of norms, as in the usual type of psychological test. Interpretive scorer reliability is concerned with the extent to which different examiners attribute the same personality characteristics to the examinee on the basis of their interpretations of the identical record. Few adequate studies of this type of scorer reliability have been conducted with projective tests. Some investigations have revealed marked divergencies in the interpretations given by reasonably well-qualified test users. A fundamental ambiguity in such results stems from the unknown contribution of the interpreter's skill. Neither high nor low scorer reliability can be directly generalized to other scorers differing appreciably from those utilized in the particular investigation.

Attempts to measure other types of test reliability have fared equally poorly in the field of projective testing. Coefficients of *internal con-*

sistency, when computed, have usually been low. In such tests as the Rorschach, TAT, and Rosenzweig P-F Study, it has been argued that different cards or items are not comparable and hence should not be used in finding split-half reliabilities. In fact, individual items in such instruments were designed to measure different variables. Moreover, the trend of responses over successive items is often considered significant in interpreting responses. One solution, of course, would be to construct a parallel form that *is* comparable, as was done in the Holtzman Inkblot Technique. It should also be noted that insofar as responses to different cards or items are combined in arriving at a total estimate of any personality characteristic, interitem agreement is assumed. Yet empirical results have rarely supported this assumption.

Retest reliability also presents special problems. With long intervals, genuine personality changes may occur which the test should detect. With short intervals, a retest may show no more than recall of original responses. When investigators instructed their subjects to write different TAT stories on a retest, in order to determine whether the same themes would recur, most of the scored variables yielded insignificant retest correlations (Lindzey & Herman, 1955). It is also relevant to note that many scores derived from projective techniques are based on very inadequate response samples. In the case of the Rorschach, for instance, the number of responses within a given individual's protocol that fall into such categories as animal movement, human movement, shading, color, unusual detail, and the like may be so few as to yield extremely unreliable indices. Large chance variations are to be expected under such circumstances. Ratios and percentages computed with such unreliable measures are even more unstable than the individual measures themselves (Cronbach, 1949, pp. 411–412).

VALIDITY. For any test, the most fundamental question is that of validity. Many validation studies of projective tests have been concerned with concurrent criterion-related validity. Most of these have compared the performance of contrasted groups, such as occupational or diagnostic groups. As was pointed out in connection with norms, however, these groups often differ in other respects, such as age or education. Other investigations of concurrent validity have used essentially a matching technique, in which personality descriptions derived from test records are compared with descriptions or data about the same subjects taken from case histories, psychiatric interviews, or long-range behavioral records. A few studies have investigated predictive validity against such criteria as success in specialized types of training, job performance, or response to psychotherapy. There has been an increasing trend to investigate the construct validity of projective instruments by testing specific hypotheses that underlie the use and interpretation of each test.

The large majority of published validation studies on projective techniques are inconclusive because of procedural deficiencies in either experimental controls or statistical analysis, or both. This is especially true of studies concerned with the Rorschach. Some methodological deficiencies may have the effect of producing *spurious evidence of validity* where none exists. An example is the contamination of either criterion or test data. Thus, the criterion judges may have had some knowledge of the subjects' test performance. Similarly, the examiner may have obtained cues about the examinee's characteristics from conversation with him in the course of test administration, or from case history material and other nontest sources. The customary control for the latter type of contamination in validation studies is to utilize blind analysis, in which the test record is interpreted by a scorer who has had no contact with the examinee and who has no information about him other than that contained in the test protocol. Clinicians have argued, however, that blind analysis is an unnatural way to interpret projective test responses and does not correspond to the way these instruments are used in clinical practice.

Another common source of spurious validity data is failure to cross-validate (see, e.g., Kinslinger, 1966). Because of the large number of potential diagnostic signs or scorable elements that can be derived from most projective tests, it is very easy by chance alone to find a set of such signs that differentiate significantly between criterion groups. The validity of such a scoring key, however, will collapse to zero when applied to new samples.

A more subtle form of error is illustrated by stereotype accuracy. Certain descriptive statements, such as might occur in a Rorschach protocol, may apply widely to persons in general, or to young men, or to hospitalized patients, or to whatever category of subjects is sampled by the particular investigation.[1] Agreement between criterion and test data with regard to such statements would therefore yield a spurious impression of validity. Some check on this error is needed, such as a measure of the agreement between the test evaluation of one subject and the criterion evaluation of another subject in the same category. This measure would indicate the amount of spurious agreement resulting from stereotype accuracy under the conditions of the particular investigation (see, e.g., Silverman, 1959).

A common source of error, arising from reliance on clinical experience in the validation of diagnostic signs, is what Chapman (1967) labelled

---

[1] The use of such generally applicable statements will be recognized as an example of the "Barnum effect" described in Chapter 16.

"illusory validation." This phenomenon may account in part for the continued clinical use of instruments and systems of diagnostic signs for which empirical validity findings are predominantly negative. In a series of experiments to test this phenomenon, Chapman and Chapman (1967) presented college students with a set of human figure drawings, each drawing accompanied by brief descriptions of two symptoms supposedly manifested by the person who made the drawing. For example, one pair included: (1) is suspicious of other people; (2) is worried about how manly he is. The students had no previous familiarity with the Draw-a-Person Test or its rationale. The symptoms chosen were those most often linked with certain drawing characteristics by a group of clinicians previously surveyed in a mail questionnaire, and were thus part of the clinical lore associated with the use of the D-A-P. In the experiment itself, however, the symptoms were distributed at random among the drawings, and different subjects received different combinations of drawings and symptoms. Thus, the subjects' learning experience did *not* support the traditional, stereotyped associations of symptoms and drawing characteristics.

Following the learning session, subjects were given a list of symptoms and asked to describe under each the drawing characteristics they had observed to be associated with it. The results showed that subjects responded in terms of preestablished popular stereotypes, even though these associations were not supported by the data presented during their learning experience. For example, they listed atypical eyes as being associated with suspiciousness; large head with worry about intelligence; and broad shoulders with concern about manliness. Not only were the interpretations unrelated to the empirical associations which the subjects had "studied," but in other experiments these stereotyped cultural associations also proved to be resistant to change under intensive training conditions designed to establish counter associations. In other words, subjects retained their a priori expectations even when exposed to contradictory observations.

Illusory correlation is a special example of the mechanism that underlies the survival of superstitions. We tend to notice and recall whatever fits our expectations; and we tend to ignore and forget whatever is contrary to our expectations. This mechanism may actually interfere with the discovery and use of valid diagnostic signs in the course of clinical observation by clinicians who are strongly identified with a particular diagnostic system. The original research of Chapman and Chapman with the D-A-P has been corroborated by similar studies with the Rorschach (Chapman & Chapman, 1969; Golding & Rorer, 1972) and with the Incomplete Sentences Blank (Starr & Katkin, 1969).

It should be noted, on the other hand, that certain inadequacies of experimental design may have the opposite effect of *underestimating* *the validity* of a diagnostic instrument. It is widely recognized, for example, that traditional psychiatric categories, such as schizophrenia, manic-depressive psychosis, and hysteria, represent crude and unrealistic classifications of the personality disorders actually manifested by patients. Hence, if such diagnostic categories are used as the sole criterion for checking the validity of a personality test, negative results are inconclusive. Similarly, failure to predict occupational criteria may reflect no more than the examiner's ignorance of the traits required for the jobs under consideration. When such criteria are employed, it is also possible that the projective test is a valid measure of the personality traits it is designed to measure, but that these traits are irrelevant to success in the chosen criterion situations.

Those who stress the importance of configural scoring, response patterns, and trait interrelationships in personality assessment have also objected to attempts to validate isolated scores or diagnostic signs derived from projective techniques. That insignificant correlations may result from failure to allow for complex patterns of relationship among personality variables can be illustrated by studies of aggression indicators in TAT stories. In interpreting the results of such investigations, it has been repeatedly pointed out that the hypothesized relation between aggression in fantasy, as revealed in the TAT, and aggression in overt behavior is not a simple one. Depending on other concomitant personality characteristics, high aggression in fantasy may be associated with either high or low overt aggression. There is some evidence suggesting that, if strong aggressive tendencies are accompanied by high anxiety or fear of punishment, expressions of aggression will tend to be high in fantasy and low in overt behavior; when anxiety and fear of punishment are low, high fantasy aggression is associated with high overt aggression (Harrison, 1965; Mussen & Naylor, 1954; Pittluck, 1950).

Lack of significant correlation between expressions of aggression in TAT stories and in overt behavior in a random sample of cases is thus consistent with expectation, since the relation may be positive in some individuals and negative in others. Obviously, however, such a lack of correlation is also consistent with the hypothesis that the test has no validity at all in detecting aggressive tendencies. What is needed, of course, is more studies using complex experimental designs that permit an analysis of the conditions under which each assumption is applicable.

Few research projects have been so designed as to avoid all the major pitfalls of projective test validation. The reader is urged to examine for himself the reports of such studies as those of Golden (1964), Henry and Farley (1959), Little and Shneidman (1959), and Silverman (1959), which set a high standard in experimental design. While differing in type of subject examined and in specific problems they set out to investigate, these studies point to a common conclusion: when experienced clinicians are given an opportunity to examine and interpret in their own way subjects' protocols from such projective tests as the Rorschach and TAT, their evaluations of the subjects' personalities tend to match independent case history evaluations significantly better than chance. Insofar as can be ascertained, however, the obtained relations are low. Moreover, the relationship appears to be a function of the particular clinician and subject, a number of individual matches being no better than chance. There is also little agreement among evaluations based on different projective techniques, or among different clinicians using the same technique.

THE PROJECTIVE HYPOTHESIS. It is a fundamental assumption of all projective techniques that the individual's responses to the ambiguous stimuli presented to him reflect significant and relatively enduring personality attributes. Yet there is a large and growing body of research data indicating that many other factors affect the individual's projective test responses. To the extent that retest reliability has been measured, marked temporal shifts have frequently been observed, indicating the operation of considerable chance error. More direct evidence regarding the susceptibility of projective test responses to temporary states is provided by several experimental studies demonstrating the effect of such factors as hunger, sleep deprivation, drugs, anxiety, and frustration on such responses.

The response variability associated with even slight changes in stimulus characteristics, furthermore, suggests that the responses are stimulus-specific and hence of questionable generalizability. Significant response differences have likewise been found in relation to instructional sets, examiner characteristics, and the examinee's perception of the testing situation. Ability factors—and particularly verbal ability—clearly affect scores on most projective tests. In the light of all these findings, projective test responses can be meaningfully interpreted only when the examiner has extensive information about the circumstances under which they were obtained and the aptitudes and experiential background of the examinee.

From another angle, the advantages of using unstructured or ambiguous stimuli have been questioned (Epstein, 1966). Such stimuli are ambiguous for the examiner as well as for the examinee; thus, they tend to increase the ambiguity of the examiner's interpretations of the subject's responses. With structured stimuli, on the other hand, it is possible to select stimuli relevant to the personality characteristics to be assessed and to vary the nature of the stimuli in a systematic and balanced manner so as to fully explore a given personality dimension. Such a procedure makes for clearer interpretation of test performance than is possible with the shotgun approach of unstructured stimuli. There is also evidence against the common assumption that the less structured the stimuli the more likely they are to elicit projection and to tap "deep" layers of personality (Murstein, 1963). Actually the relation between ambiguity and projection appears to be nonlinear, with an intermediate degree of ambiguity representing an optimum for purposes of projection.

The assumption that fantasy, as elicited by such projective techniques as the TAT, reveals covert motivational dispositions has also been called into question. One survey of the experimental literature on the projective assessment of achievement need indicated that the data were more nearly consistent with alternative explanations involving perceptual and cognitive factors (Klinger, 1966). Similarly, in a 20-year longitudinal study of TAT fantasy and relevant overt behavior, adolescent activities predicted adult TAT imagery much better than adolescent TAT imagery predicted adult activities (McClelland, 1966; Skolnick, 1966). For example, individuals who had shown upward social mobility obtained higher scores on achievement needs as adults; but those who obtained higher achievement need scores in adolescence were not among the ones who subsequently showed upward social mobility.

Findings such as these reverse the relationship implied by the traditional rationale of projective techniques. They can be explained if we regard TAT responses, not as direct projective expressions of motives, but as samples of the individual's thoughts, which may in turn have been influenced by his previous actions. Individuals who have achieved more and those who were more often exposed to achievement-oriented models in their developmental history tend to perceive more achievement themes in unstructured pictures.

In summary, many types of research have tended to cast doubt on the projective hypothesis. There is ample evidence that alternative explanations may account as well or better for the individual's responses to unstructured test simuli.

PROJECTIVE TECHNIQUES AS PSYCHOMETRIC INSTRUMENTS. Besides their questionable theoretical rationale, projective techniques are clearly found wanting when evaluated in accordance with test standards. This is evident from the data summarized in the preceding sections with regard to standardization of administration and scoring procedures, adequacy of norms, reliability, and validity. The accumulation of published studies that have *failed* to demonstrate any validity for such projective techniques as the Rorschach and the D-A-P is truly impressive. Yet after four decades of negative results, the status of projective techniques remains substantially unchanged. In the words of one reviewer, "There are still enthusiastic clinicians and doubting statisticians" (Adcock, 1965).

This apparent contradiction can perhaps be understood if we recognize that, with a few exceptions, projective techniques are not truly tests.

Among the notable exceptions are the Holtzman Inkblot Technique, the Adkins and Ballif Animal Crackers test, and the Rosenzweig P-F Study. It is interesting to note in this connection that these instruments have proved to be equally applicable in group and individual forms—a finding not generally characteristic of projective techniques, where the interpersonal relation between examiner and examinee is often an integral part of the examining process. Even in the case of these instruments, however, there is need for much more validity data to specify the nature of the constructs measured by their scores, as well as for more normative data on clearly defined populations. Thus, while coming closer to meeting test standards than have other projective techniques, even these instruments are not ready for routine operational use.

A few other examples of quasi-tests could undoubtedly be found among the many remaining projective techniques that were not discussed in this chapter. Such exceptions, however, only serve to demonstrate that unstructured stimuli may be used in constructing psychometric instruments. The fact remains, nevertheless, that they have not been so used in the vast majority of projective techniques, particularly in the more popular and firmly entrenched techniques.

PROJECTIVE TECHNIQUES AS CLINICAL TOOLS. Rather than being regarded and evaluated as psychometric instruments, or *tests* in the strict sense of the term, projective techniques are coming more and more to be regarded as clinical tools. Thus, they may serve as supplementary qualitative interviewing aids in the hands of a skilled clinician.[2] Their value as clinical tools is proportional to the skill of the clinician and hence cannot be assessed independently of the individual clinician using them. Attempts to evaluate them in terms of the usual psychometric procedures would thus be inappropriate. But by the same token, the use of elaborate scoring systems that yield quantitative scores is not only wasteful but also misleading. Such scoring procedures lend the scores an illusory semblance of objectivity and may create the unwarranted impression that the given technique can be treated as a test. What value projective techniques may have is more likely to emerge when they are interpreted by qualitative, clinical procedures than when they are quantitatively scored and interpreted as psychometric instruments.

Borrowing a concept from information theory, Cronbach and Gleser (1965) characterized interviewing and projective techniques as "wide-band" procedures. Bandwidth, or breadth of coverage, is achieved at the cost of lowered fidelity or dependability of information. Objective psychometric tests characteristically yield a narrow band of information at a high level of dependability. In contrast, projective and interviewing techniques provide a much wider range of information of lower dependability. Moreover, the kinds of data furnished by any one projective technique may vary from individual to individual. One person's TAT responses, for example, may tell us a good deal about his aggression and little or nothing about his creativity or achievement need; another person's record may permit a thorough assessment of the degree of creativity of his behavior and of the strength of his achievement need, while revealing little about his aggression. Such a lack of uniformity in the kinds of information provided in individual cases helps to explain the low validities found when projective test responses are analyzed for any single trait across a group of persons.

---

[2] A stimulating account by Kaplan (1975) provides examples of the use of both projective and standardized instruments in nonstandard ways in the clinical examination of otherwise untestable children.

It is interesting to note that a similar unevenness characterizes clinicians' interpretations of individual records. Thus, in their study of the validity of the TAT, Henry and Farley (1959, p. 22) concluded:

> There is no single correct way of employing the TAT interpretation. There was little item agreement between judges, but each judge made enough "correct" decisions to yield a highly significant agreement figure. Judges may arrive at essentially the same interpretive implications of the test report, by quite different routes; or judges may differ individually in their abilty to utilize TAT predictions in different areas . . . or for different subjects.

The nature of clinical judgment through which projective and interviewing data may be utilized in reaching decisions about individual cases is receiving increasing attention from psychologists (see Ch. 16). In this process, the very constructs or categories in terms of which the data are organized are built up inductively through an examination of the particular combination of data available in the individual case. The special function of the clinician is to make predictions from unique or rare combinations of events about which it is impracticable to prepare any statistical table or equation. By creating new constructs to fit the individual case, the clinician can predict from combinations of events that he has never encountered before in any other case. In making his predictions, he can also take into account the varied significance of similar events for different individuals. Such clinical predictions are helpful, provided they are not accepted as final but are constantly tested against information elicited through subsequent inquiry, test responses, reaction to therapy, or other behavior on the part of the client. It follows from the nature of interviewing and projective techniques that decisions should not be based on any single datum or score obtained from such sources. These techniques serve best in sequential decisions, by suggesting leads for further exploration or hypotheses about the individual for subsequent verification.

CHAPTER 20

Other Assessment Techniques

THE SELF-REPORT inventories and projective techniques surveyed in the preceding chapters represent the best-known and most widely used instruments for personality appraisal. Nevertheless, there still remains a rich supply of other devices that are being explored for this purpose. Out of this diversity of approaches may come techniques that will eventually stimulate progress in new directions. The procedures to be considered in this chapter are principally research techniques, although some may also serve as supplementary clinical aids in the hands of a skilled clinician. A wide variety of approaches is represented by the specific assessment techniques cited. Three major categories include "objective" performance tests utilizing perceptual, cognitive, or evaluative tasks; several kinds of situational tests; and techniques designed to assess self-concepts and personal constructs. To add further perspective to this survey of personality tests, brief mention will be made of the type of assessment techniques employed in behavior modification programs; the role of direct behavioral observations and ratings in personality assessment; the analysis of life-history data for predictive purposes; and the assessment of environments.

"OBJECTIVE" PERFORMANCE TEST

A large and heterogeneous category of personality tests comprises relatively simple and objective procedures, which usually require perceptual, cognitive, or evaluative responses. The examinee is given a task that bears little resemblance to the criterion behavior under investigation. For this reason, these techniques are sometimes described as "indirect" tests.

Although varying widely in content and specific techniques, these tests have several common distinguishing features. First, the examinee is *task-oriented*, rather than being report-oriented as in personality questionnaires. He is given an objective task to perform, rather than being asked to describe his habitual behavior. Second, the purpose of these tests is *disguised*, the individual not realizing which aspects of his performance are to be scored. Third, the tasks set for the examinee are *structured*.

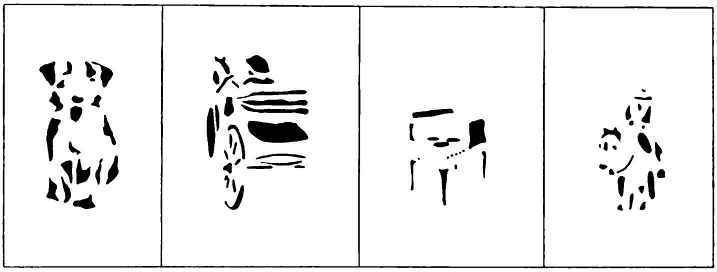
In this feature lies their principal difference from the tasks utilized in projective techniques. To be sure, structuring is a matter of degree. As a group, however, the tests now under consideration are more highly structured than are typical projective devices. A fourth and related feature pertains to the *apparent existence of a "right solution"* for each task or problem—at least from the examinee's viewpoint. Many of the tests are perceived as aptitude measures, in which the examinee endeavors to give "correct" answers. Thus, the individual's approach to the test is quite unlike that encouraged by projective tests, in which "anything goes."

Several tests in the present category were developed in the course of long-term research projects. The most extensive collection has been assembled by Cattell and Warburton (1967), as one approach to Cattell's comprehensive factor-analytic identification of personality traits (see Ch. 17). These tests are quite varied, ranging from physiological measures to tests of musical preference and the appreciation of humor.

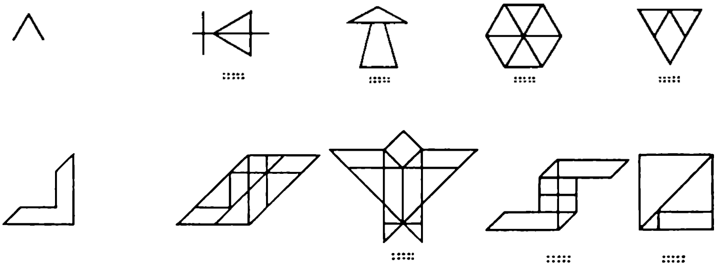
Some objective personality tests have been constructed as measures of cognitive styles, as illustrated by the research of Witkin and his associates (Witkin et al., 1954, 1974). Cognitive styles refer essentially to one's preferred and typical modes of perceiving, remembering, thinking, and problem solving (Messick, 1970, 1973b). They are regarded as broad stylistic behavioral characteristics that cut across abilities and personality and are manifested in many activities and media.

PERCEPTUAL FUNCTIONS. One of the principal sources of the simple, objective personality tests under consideration is to be found in the area of perceptual functions. A large body of experimental literature has demonstrated significant relationships between the individual's attitudinal, motivational, or emotional characteristics and his performance on perceptual or cognitive tasks. It should also be recognized that a number of projective techniques—notably the Rorschach—are essentially perceptual tests.

Of the factors identified in factorial analyses of perception, two that have proved particularly fruitful in personality research are speed of closure and flexibility of closure (Pemberton, 1952; Thurstone, 1944). The first involves the rapid recognition of a familiar word, object, or other figure in a relatively unorganized or mutilated visual field. Typical items from a test found to be highly saturated with this factor (Street Gestalt Completion) are reproduced in Figure 105. Flexibility of closure requires the identification of a figure amid distracting and confusing details. Two items from a test with a high loading in this factor (Gottschaldt Figures) are also shown in Figure 105. Several studies have reported suggestive data indicating possible relationships between each of these factors and personality traits. In one investigation (Pemberton,



Speed of Closure (Street Gestalt Completion): What does each picture represent?



Flexibility of Closure (Gottschaldt Figures): Which of the four drawings at the right contain the design at the left?

Fig. 105. Sample Items Illustrating Perceptual Tests.
(From Thurstone, 1950, p. 7.)

1952), for example, persons who excelled in speed of closure tended to rate themselves as sociable, quick in reactions, artistic, self-confident, systematic, neat and precise, and disliking logical and theoretical problems. In contrast, those scoring high in flexibility of closure had high self-ratings on such traits as socially retiring, independent of the opinions of others, analytical, interested in theoretical and scientific problems, and disliking rigid systematization and routine. Commercially available forms of these two perceptual tests have been published under the titles of Closure Speed and Closure Flexibility.

Considerable research has been done on the individual's ability to resist the disruptive influence of conflicting contextual cues in his perceptions of visual forms and relationships. Attention was first focused on this perceptual variable by the work of Witkin and his associates (1954) on perceptual space orientation. Through various tests utilizing a rod and frame that could be independently moved, a tilting chair, and a tilting room, these investigators were able to show that individuals differ widely in their "field dependence," or the extent to which their perception of the upright is influenced by the surrounding visual field.

A substantial body of data was amassed to indicate that field dependence is a relatively stable, consistent trait, having a certain amount of generality. Thus, both odd-even and retest reliabilities were high, and most of the intercorrelations among the different spatial orientation tests were significant. Of even more interest are the significant correlations between these orientation tests and the Embedded Figures Test (similar to the Gottschaldt Figures illustrated in Figure 105), which may be regarded as measuring field dependence in a purely visual, paper-and-pencil situation. As more research accumulated, field dependence came to be regarded as the perceptual component of a broader personality dimension, designated as global versus articulated cognitive style, or psychological differentiation (Witkin, Dyk, Faterson, Goodenough, & Karp, 1974). There is evidence that this cognitive style exhibits considerable stability through childhood and early adulthood and is related to a number of personality variables, such as leadership (Weissenberg & Gruenfeld, 1966) and social conformity (Witkin et al., 1974). Most of this research has used the Embedded Figures Test, which is relatively easy to administer. Published forms of this test are now available for the adult, child, and preschool level, as well as for group administration.

AESTHETIC PREFERENCES. Another approach utilizes aesthetic preferences as a means of assessing personality styles. This technique is illustrated by the Barron-Welsh Art Scale, in which the examinee simply records whether he likes or dislikes each of 86 black and white figures. Two examples are reproduced in Figure 106. The items in this scale were originally selected from a pool of 400 figures because they differentiated between the responses of contrasted criterion groups of artists and nonartists. Subsequent research suggested, however, that performance on the scale may be related to creativity, not only in art but also in other fields (Barron, 1965).

There are also considerable data on the relation of the Barron–Welsh scale to personality variables. For example, low scorers on this scale tended more often to yield to social pressure, while high scorers tended more often to form independent judgments. In self-descriptions, high scorers more often chose such terms as adventurous, argumentative,

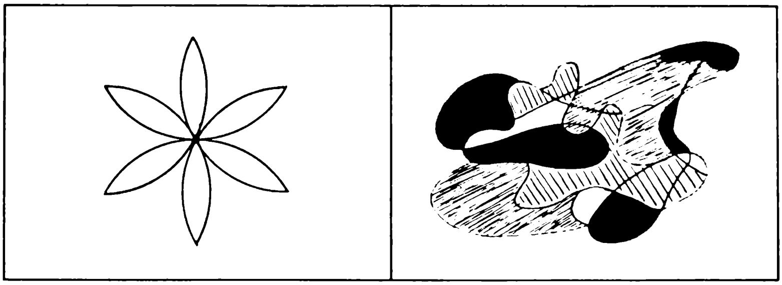


FIG. 106. Typical Figures from the Barron-Welsh Art Scale. For each figure, the respondent merely records "Like" or "Don't Like."

(Reproduced by permission of George S. Welsh.)

dominant, impulsive, shrewd, and unconventional, while low scorers more often chose the adjectives conservative, conventional, industrious, persevering, and thorough, among others. This test has been used extensively in research and has shown some promise as an indicator of fairly broad personality variables.

HUMOR. Reactions to humor have likewise been explored as possible indicators of personality variables. The IPAT Humor Test of Personality (Cattell & Luborsky, 1947; Luborsky & Cattell, 1947), available in two forms, provides jokes and cartoons to be evaluated. Form A consists of 104 pairs of jokes, in each of which the respondent chooses the joke he considers funnier. A sample item is shown below:

(*a*) Epitaph to a waiter: By and by God caught his eye.	(*b*) One prehistoric man to another: "Now that we've learned to communicate with each other—shut up!"

In Form B, 130 jokes or cartoons are individually rated as funny or dull. Although Form A controls the response set to check many or few jokes as funny, Form B provides additional information about the individual's general tendency to regard jokes as funny. For scoring purposes, the items on the Humor test were grouped into clusters on the basis of factor analysis. A separate score is reported for each of 13 personality factors.

EVALUATION OF PROVERBS. Still another technique utilizes reactions to proverbs or aphorisms. In the Famous Sayings test (Bass, 1956), the examinee responds to each of 130 statements by indicating whether he agrees, disagrees, or is uncertain. Scales identified as Hostility, Fear of Failure, and Conventional Mores were derived through factor analysis.

A Social Acquiescence score is also found on the basis of the respondent's general acceptance or rejection of items. Apart from the factorial analyses and a few significant correlations with certain personality inventory scores, validity data for this test were based largely on significant differences between various occupational, regional, educational, and clinical groups (Bass, 1957; T. M. Walsh, 1966).

INTERESTS AND ATTITUDES. Finally, mention may be made of the use of relatively objective, task-oriented tests in the appraisal of interests and attitudes (D. T. Campbell, 1950, 1957). Several information tests have been developed for research on attitudes. The knowledge an individual has acquired is apt to reflect his selective perception and retention of facts, as well as his biased sources of information. We are likely to notice and to remember those facts that are in line with our expectations or hypotheses, and to overlook and forget others. Moreover, when no information is available on a given question, the direction of guessing may be determined by the respondent's attitude; this tendency has been utilized in the *error-choice technique* (D. T. Campbell, 1957), in which the respondent is forced to choose between two equally incorrect alternatives reflecting opposed biases. The items in error-choice tests are such that the correct answers are not generally familiar to the majority of respondents. Hence, the errors are not readily apparent. Attitude bias on the part of the individual is revealed by systematic errors in one direction, as opposed to random errors.

Among the many other techniques utilized for the measurement of attitudes may be mentioned perception and memory tests, in which distortions and errors reflect bias; the evaluation of arguments, syllogistic conclusions, inferences, and the like; prediction of the outcomes of described events; estimation of group opinions, the estimated opinions being presumably colored by the individual's own views; responding in various ways to photographs of persons identified as members of different minority groups, occupations, or other categories; and rating jokes that pertain to contrasting orientations, viewpoints, or other group characteristics (e.g., LaFave, Haddad, & Marshall, 1974).

SITUATIONAL TESTS

Although the term "situational test" was popularized during and following World War II, tests fitting this description had been developed prior to that time. Essentially, a situational test is one that places the examinee in a situation closely resembling or simulating a "real-life" criterion situation. Such tests thus show certain basic similarities to the job-sample technique employed in constructing vocational achievement tests (Ch. 15). In the present tests, however, the criterion behavior that is sampled is more varied and complex. Moreover, performance is not evaluated in terms of abilities and knowledge, but in terms of emotional, social, attitudinal, and other personality variables.

TESTS OF THE CHARACTER EDUCATION INQUIRY. Among the earliest situational tests—although they were not so labeled at the time—were those constructed by Hartshorne, May, and their associates (1928, 1929, 1930) for the Character Education Inquiry (CEI). These tests were designed principally as research instruments for use in an extensive project on the nature and development of character in children. Nevertheless, the techniques can be adapted to other testing purposes, and a number have been so utilized.

In general, the CEI techniques made use of familiar, natural situations within the schoolchild's daily routine. The tests were administered in the form of regular classroom examinations, as part of the pupil's homework, in the course of athletic contests, or as party games. Moreover, the children were not aware that they were being tested, except insofar as ordinary school examinations might be involved in the procedure. At the same time, all of the Hartshorne-May tests represented carefully standardized instruments which yielded objective, quantitative scores.

The CEI tests were designed to measure such behavior characteristics as honesty, self-control, and altruism. The largest number were concerned with honesty. One group of tests developed for this purpose utilized a duplicating technique. Common tests such as vocabulary, sentence completion, or arithmetic reasoning were administered in the classroom. The test papers were then collected and a duplicate of each child's responses was made. At a subsequent session, the original, unmarked test papers were returned, and each child scored his own paper from a key. Comparison with the duplicate record revealed any changes the individual had made in scoring his paper. Another technique was based on improbable achievement. In the Circles Puzzle, for example, the child was instructed to make a mark in each of 10 small, irregularly arranged circles, while keeping his eyes shut. Control tests under conditions that precluded peeking indicated that a score of more than 13 correctly placed marks in a total of three trials was highly improbable. A score above 13 was therefore recorded as evidence of peeking.

Self-control was assessed in part by the length of time the child persisted in a given task, such as solving a difficult puzzle or reading the ending of a story in which the words were run together. In the latter test, the examiner read aloud the beginning of an exciting story. The children were then given pages containing the rest of the story printed in a form that made reading difficult. As the end of the story was approached, moreover, the difficulty of deciphering the words increased further, as shown in the three successive samples given below (Hartshorne, May, & Maller, 1929):

> CHARLESLIFTEDLUCILLETOHISBACK"PUTYOURARMSTIGHT
> AROUNDMYNECKANDHOLDON
>
> NoWhoWTogETBaCkONthETREStle.HoWTOBRingTHaTTerrIfIED
> BURDeNOFAChILDuPtOSafeTY
>
> fiN ALly tAp-tAPC AME ARHYTH Month eBriD GeruNNing fee Tfee
> Tcom INGtow ArdT Hem

Most of the CEI tests proved to have good discriminative power, yielding a wide range of individual differences in scores. Reliability also appeared to be fairly satisfactory. The children's responses, however, showed considerable situational specificity. The intercorrelations of different tests within each category (e.g., honesty tests or persistence tests) proved to be very low. This specificity is understandable when we consider the operations of the child's interests, values, and motives in different situations. For example, the child who is motivated to excel in school work is not necessarily concerned about his achievement in athletic contests or party games. These motivational differences would in turn be reflected in the child's behavior on honesty tests administered in these different contexts. The influence of individual teachers was also apparent in the results; children who cheated in one teacher's room would not do so in another teacher's room. Both in their findings and in their interpretations, the CEI investigators anticipated the present emphasis on situational specificity by four decades. A recent reanalysis of the CEI data, by sophisticated statistical techniques, suggests that a more appropriate model includes both a general factor of honesty and a situational component (Burton, 1963).

SITUATIONAL STRESS TESTS. Like the tests developed in the CEI program, situational stress tests are realistic and disguised. The principal difference in the case of stress tests is to be found in the introduction of features designed to induce anxiety or other disruptive emotional states. The individual's reactions to stress may be assessed through objective records of performance on the assigned tasks, qualitative observations by trained observers, or both. Among the most ambitious projects concerned with the development and use of situational stress tests was that undertaken by the United States Office of Strategic Services (OSS) during World War II (Murray & MacKinnon, 1946; OSS Assessment Staff, 1948; J. S. Wiggins, 1973, Ch. 11). The object of this testing program was the evaluation of candidates for assignment to military intelligence. The principal assessment program consisted of a 3-day session of intensive testing and observation. During this period, the candidates lived together in small groups under almost continuous scrutiny by members of the assessment staff. Besides specially constructed situational tests, the program included aptitude tests, projective techniques, intensive interviewing, and general observations under casual and informal conditions. Several of the situational tests were modeled after techniques developed in the German and British armies.

An example of an OSS situational stress test is provided by the Construction Test, in which a 5-foot cube had to be assembled from wooden poles, blocks, and pegs. The examinee was informed that, since it was impossible for one man to complete the task within the 10 minutes allotted to it, he would be given two helpers. Actually, the helpers were psychologists who played prearranged roles. One followed a policy of inertia and passive resistance; the other obstructed the work by making impractical suggestions, asking irrelevant and often embarrassing questions, and needling the candidate with ridicule and criticism. So well did the helpers succeed in frustrating the candidates that the construction was never completed in the history of the assessment program. The subject's emotional and interpersonal reactions in this situation were observed and evaluated qualitatively.

Available validity data on situational stress tests are meager (Guilford & Lacey, 1947; E. L. Kelly, 1954; OSS Assessment Staff, 1948; J. S. Wiggins, 1973, Ch. 11). A major difficulty arises from the lack of suitable criterion data. In the OSS project, for example, the diversity of individual assignments reduced the possibility of evaluating field performance under comparable conditions for different members of the group. Unreliability of criterion ratings also tends to reduce validity estimates. With these qualifications, it must be noted that, when determined, predictive validity of situational stress tests has proved to be low. Although a few techniques may be promising for special purposes, the contribution of these tests rarely justifies the time, equipment, and highly trained personnel that their administration requires.

LEADERLESS GROUP SITUATIONS. A relatively common type of situational test utilizes a "leaderless group" as a device for appraising such characteristics as cooperation, teamwork, resourcefulness, initiative, and leadership. In such tests, a task is assigned that requires the cooperative efforts of a group of examinees, none of whom is designated as leader or given specific responsibilities. Examples from the OSS program include the Brook Situation, involving the transfer of personnel and equipment across a brook with maximum speed and safety; and the Wall Situation, in which men and materials had to be conveyed over a double wall separated by an imaginary canyon.

A promising variant of this technique is the Leaderless Group Discussion (LGD). Requiring a minimum of equipment and time, this technique has been used widely in the selection of such groups as military officers, civil service supervisors and administrators, industrial executives and management trainees, sales trainees, teachers, and social workers (Bass, 1954). It has also been employed in research on leadership (Bass, 1959; Carter, 1954), on the effects of counseling (Pepinsky, Siegel, & Vanatta, 1952), and on the selection of clinical psychology trainees (Kelly & Fiske, 1951). Essentially, the group is assigned a topic for discussion during a specified period. Examiners observe and rate each person's performance, but do not participate in the discussion. Although often used under informal and unstandardized conditions, the LGD has been subjected to considerable research. Rater reliability has been found to be reasonably high for both LGD and other situational tests based on group decision-making through discussion (Greenwood & McNamara, 1967).

Validity studies suggest that LGD techniques are among the most effective applications of situational tests. Many significant and sizeable correlations have been found between ratings from LGD performance and follow-up or concurrent ratings obtained in military, industrial, and social settings (Bass, 1954; Guilford, 1959). Some of these correlations are as high as .60. It is also interesting to note that leadership ratings based on a one-hour LGD correlated over .60 with leadership assessment from three days of situational testing in the OSS program (OSS Assessment Staff, 1948). A similar correlation between LGD and an entire battery of situational tests was found by Vernon (1950) in a British study.

Neither LGD nor other, more elaborate situational tests, however, have proved valid as devices for assessing broad personality traits (E. L. Kelly, 1954; Kelly & Fiske, 1951; OSS Assessment Staff, 1948). All such tests appear to be most effective when they approximate actual work-samples of the criterion behavior they are designed to predict. The LGD tests in particular have some validity in predicting performance in jobs requiring a certain amount of verbal communication, verbal problem-solving, and acceptance by peers. Another factor that seems to increase the predictive validity of situational tests is job familiarity on the part of the raters (Holmen et al., 1956; Vernon, 1950). Such a finding again suggests that situational tests work best when the examinee's performance is interpreted as a worksample rather than in terms of underlying personality variables.

ROLE PLAYING. Some situational tests employ role playing to obtain a sample of interpersonal, job-relevant behavior. The problem situations presented in these tests involve the sort of face-to-face contacts common in administrative jobs. Several such tests have been developed for military use in assessing the leadership and human-relations skills of officers (Flanagan, 1954b). An example of such a role-playing situation is described below (American Institutes for Research, 1957):

> The examinee is evaluated as the commanding officer of a small ship. He must discuss with Ensign Baker, the ship's engineering officer, an unsatisfactory performance report he is about to submit. Although Baker is highly qualified in the technical skills required, he lacks experience. He has allowed the enlisted men to run his department. This has led to no specific incident in which his department has interfered with the ship's operating efficiency, but his behavior has set a bad example for the men and the other officers on board.

The examinee is instructed to assume the given role and carry out the interview with "Ensign Baker," whose part is played by a trained actor. The examiner observes and evaluates the individual's performance with the aid of a previously developed checklist of effective and ineffective behaviors. Examples of such behaviors include: Gives reason for meeting; allows Baker to direct conversation; vacillates in, or changes position on, unsatisfactory report; mentions Baker's fine engineering background; mentions constructive nature of Fitness Report.

SELF-CONCEPTS AND PERSONAL CONSTRUCTS

PHENOMENOLOGICAL INFLUENCE. A number of current approaches to personality assessment concentrate on the way the individual views himself and others. Such techniques reflect the influence of phenomenological psychology, which focuses on how events are *perceived* by the individual (G. A. Kelly, 1963, 1970; Snygg & Combs, 1959). The individual's self-description thus becomes of primary importance in its own right, rather than being regarded as a second-best substitute for other behavioral observations. Interest also centers on the extent of self-acceptance shown by the individual.

It might be argued that self-concept tests do not differ essentially from the self-report inventories discussed in Chapter 17. True, but it would be more accurate to say that self-report inventories are actually measures of self-concept. The interpretation of personality inventory responses in terms of self-conceptualization forms the basis of a provocative hypothesis formulated by Loevinger (1966a, 1966b; Loevinger & Ossorio, 1958; Loevinger, Wessler, & Redmore, 1970). Bringing together many disparate findings from her own research and that of others, Loevinger proposed a personality trait which she defined as the capacity to conceptualize oneself, or to "assume distance" from oneself and one's impulses. According to Loevinger, it is the manifestations of this trait in personality inventories that have been described in such terms as façade, test-taking defensiveness, response set, social desirability, acquiescence, and personal style. In common with a number of other psychologists, Loevinger regards such test-taking attitudes, not as instrumental errors to be ruled out, but as the major source of valid variance in personality inventories.

On the basis of data from many sources, Loevinger suggested that ability to form a self-concept increases with age, intelligence, education, and socioeconomic level. At the lowest point, illustrated by the infant, the individual is incapable of self-conceptualization. As the ability develops, he gradually forms a stereotyped, conventional, and socially acceptable concept of himself. This stage Loevinger considers to be typical of adolescence. With increasing maturity, the individual progresses beyond such a stereotyped concept to a differentiated and realistic self-concept. At this point, he is fully aware of his idiosyncrasies and accepts himself for what he is. It is this trait of self-conceptualization, later designated as ego development or I-level, that Loevinger and her associates undertook to measure in the Washington University Sentence Completion Test, described in Chapter 19.

According to Loevinger, many (if not most) persons fail to reach the final stage of differentiated self-concept. Insofar as personality inventory responses are evaluated in terms of normative data, individuals whose self-concepts are at the stereotyped conventional stage receive higher or "better adjusted" scores. In the course of psychotherapy, some persons may advance beyond this stage to the individualized self-concept and hence may show a decline in scores on adjustment inventories (Loevinger & Ossorio, 1958). Such a hypothesis could account for the apparent failures of personality inventories when used in a clinical setting. The finding by some investigators (Sanford, 1956) that, when evaluated in terms of personality inventory norms, college seniors appear to have poorer emotional adjustment than college freshmen may have a similar explanation. Essentially, Loevinger argues that the capacity for self-conceptualization is an important personality trait and that the relation of this trait to personality inventory scores is not linear but curvilinear.

The procedures to be considered in this section all have as their primary focus the individual's perception of himself and others. Another common feature of these procedures is their applicability to intensive investigation of the individual case. For this reason, they are of special interest to the clinical psychologist. Many of them, in fact, have originated within a clinical setting.

THE ADJECTIVE CHECK LIST. Several techniques have been specifically developed for assessing self-concepts (Strong & Feder, 1961). A widely applicable instrument that is commercially available is the Adjective Check List (ACL). Originally constructed for use in the research program of the Institute for Personality Assessment and Research (IPAR), this instrument provides a list of 300 adjectives arranged alphabetically from "absent-minded" to "zany" (Gough, 1960; Gough & Heilbrun, 1965). The respondent marks all the adjectives he considers to be descriptive of himself.

The ACL can be scored for 24 basic scales. Three are response set scales, including the total number of words marked and the number of these that are favorable and unfavorable, respectively. The total number of items marked is not only used to adjust the scores on the other scales but is also of interest in its own right. Persons marking many adjectives tend to be described as active and enthusiastic; those marking few adjectives, as quiet, reserved, and cautious. Four scales were developed by empirical item keying against a criterion of intensive personal observation and evaluation of participants in the IPAR assessment program; these include the scales of Self-confidence, Self-control, Lability,[1] and Personal adjustment. Another empirical scale, labeled Counseling readiness, was developed by comparing the responses of contrasting groups of counseling clients, one showing more positive and the other less positive response to counseling. The remaining 15 scales were prepared on a rational or content basis, by assigning adjectives to each of the 15 needs covered by the EPPS (see Ch. 17). Examples include the needs for Achievement, Dominance, Affiliation, Order, and Autonomy. The ACL is thus one more instrument reflecting the influence of Murray's classification of personality needs.

Although assessments of participants in the IPAR program were used in developing only four of the scales, they were subsequently employed in building up the construct validity of other scales as well. On the basis of these assessments as well as other research, the manual provides personality sketches of individuals scoring high and low on each scale. While considerable information is available about some of these scales, the manual properly characterizes all the scales as being experimental. As a research instrument, however, the ACL has been applied to a staggering variety of problems, drawn from such areas as psychopathology, occupational choice, creativity, political and economic behavior, and even patients' reactions to orthodontia and to contact lenses. It has also been employed in rating historical personages from their biographies and published works (Welsh, 1975) and in characterizing inanimate objects, such as cities or automobiles.[2]

[1] Refers to spontaneity, flexibility, need for change, rejection of convention, and assertive individuality.

[2] A complete ACL bibliography is now included in the test manual. A general review of the adjective checklist technique is given by Masterson (1975).

Q SORT. Another special technique suitable for investigating self-concepts is the *Q* sort originally developed by Stephenson (1953). In this technique, the individual is given a set of cards containing statements or trait names which he must sort into piles ranging from "most characteristic" to "least characteristic" of himself. The items may come from a standard list, but more often are designed to fit the individual case. To ensure uniform distribution of ratings, a "forced-normal" distribution is used, the respondent being instructed to place a specified number of cards in each pile. Such a distribution can be prepared for any size of item sample by reference to a normal curve table. It should be noted that, like the forced-choice technique discussed in Chapter 17, the *Q* sort yields ipsative rather than normative data. In other words, the individual tells us which he considers his strong and which his weak traits, but not how strong he believes himself to be in comparison with another person or some outside norm.

Q sorts have been employed to study a variety of psychological problems (Block, 1961, 1971a; S. R. Brown, 1968; Mowrer, 1953; Rogers & Dymond, 1954). In intensive investigations of individual personality, the subject is often asked to re-sort the same set of items within different frames of reference. For example, he may sort the items as they apply to himself and to other persons, such as his father, his mother, or his wife. Similarly, he may sort the items as they apply to himself in different settings, such as job, home, or social situations. *Q* sorts can likewise be obtained for the individual as he believes he actually is (real self), as he believes others see him (social self), and as he would like to be (ideal self). To observe change, *Q* sorts may be obtained successively at different stages during psychotherapy, a procedure that has been followed especially by client-centered therapists. With therapy, the self-concept tends to become more favorable and to resemble more closely the individual's ideal-self concept[3] (Rogers & Dymond, 1954, Ch. 4).

THE SEMANTIC DIFFERENTIAL. This technique was first developed by Osgood and his associates (1957) as a tool for research on the psychology of meaning, although its possibilities for personality assessment were soon recognized. The Semantic Differential represents a standardized and quantified procedure for measuring the connotations of any given concept for the individual. Each concept is rated on a 7-point graphic scale as being more closely related to one or the other of a pair of opposites, as illustrated in Figure 107. For every concept, a series of these bipolar

[3] All these procedures can, of course, be followed also with the previously described Adjective Check List or with any other techniques for assessing self-concepts.

adjectival scales is employed; usually 15 or more scales are included. Intercorrelations and factorial analyses of the original set of 50 scales developed by Osgood revealed three major factors: *Evaluative*, with high loadings in such scales as good-bad, valuable-worthless, and clean-dirty; *Potency*, found in such scales as strong-weak, large-small, and heavy-light; and *Activity*, identified in such scales as active-passive, fast-slow, and sharp-dull. The evaluative factor is the most conspicuous, accounting for the largest percentage of total variance.

FATHER

Good ____:____:____:____:____:____:____ Bad
Clean ____:____:____:____:____:____:____ Dirty
Cruel ____:____:____:____:____:____:____ Kind
Slow ____:____:____:____:____:____:____ Fast
Valuable ____:____:____:____:____:____:____ Worthless
Tense ____:____:____:____:____:____:____ Relaxed
Strong ____:____:____:____:____:____:____ Weak
Large ____:____:____:____:____:____:____ Small

FIG. 107. Illustration of the Semantic Differential Technique. In rating the concept "Father," respondent places a check mark on the appropriate segment of each scale. Usually a much larger number of scales is employed.

Responses on the Semantic Differential can be analyzed in several ways. For quantitative treatment, the ratings on each scale can be assigned numerical values from 1 to 7, or from −3 to +3. The overall similarity of any two concepts for an individual or a group can then be measured in terms of their positions on all scales. The connotations of all concepts rated by a single individual can be investigated by computing the "score" of each concept in the three principal factors described above. Thus, on a scale extending from −3 to +3, a given individual's concept of "My Brother" may rate −2 in the evaluative factor, 0.1 in potency, and 2.7 in activity.

The concepts to be rated can be chosen to fit whatever problem is being investigated. The respondent can, for example, be asked to rate: himself, members of his family, friends, employers, teachers, or public figures; members of different ethnic or cultural groups; persons engaged in different occupations; activities, such as studying or outdoor sports; abstract ideas, such as confusion, hatred, sickness, peace, or love; product names or brand names; and radio or television programs. The Semantic Differential has been applied in many different contexts, in research on such varied problems as clinical diagnosis and therapy, vocational choices, cultural differences, and consumers' reactions to products and brand names (Snider & Osgood, 1969). The bibliography on the Semantic Differential includes some 1,500 references.

ROLE CONSTRUCT REPERTORY TEST. A technique devised specifically as an aid in clinical practice is the Role Construct Repertory Test (Rep Test) developed by G. A. Kelly (1955, 1963, 1970). This technique shares certain features with other testing procedures, notably the Semantic Differential and the various sorting tests used to study concept formation (Ch. 16). In the Rep Test, however, the objects to be sorted are persons who are important in the individual's life. And unlike the Semantic Differential, the Rep Test requires the individual himself to designate the scales, dimensions, or constructs in terms of which he characterizes these persons.

The development of the Rep Test is intimately related to Kelly's personality theory. A basic proposition in this theory is that the concepts or constructs an individual uses to perceive objects or events influence his behavior. In the course of psychotherapy, it is frequently necessary to build new constructs and to discard some old constructs before progress can be made.

The Rep Test is designed to help the clinician identify some of the client's important constructs about people. Although the test can be administered in many ways, including both group and individual versions, one of its simpler variants will serve to illustrate its essential characteristics.[4] In this variant, the examinee is first given a Role Title List and asked to name a person in his experience who fits each role title. Typical roles might include: your father, your wife or present girlfriend, a teacher you liked, a person with whom you have been closely associated recently who appears to dislike you. The examiner next selects three of the persons named and asks, "In what *important way* are two of them alike but different from the third?" This procedure is repeated with many other sets of three names, in which some of the names recur in different combinations.

The Rep Test yields a wealth of qualitative data. A simplified factor-analytic procedure has also been developed for quantitative identification of constructs that are important for each individual. Other investigators have identified major dimensions of role constructs through analyses of group data (Messick & Kogan, 1966). When individuals were scored on these dimensions, the scores were found to correlate significantly with certain cognitive and personality variables. In another study, retests over a 2-week interval with both identical and different instructions revealed a high level of stability in the constructs employed by individuals (Fjeld & Landfield, 1961).

---

[4] For descriptions of other procedural variants, see Bannister and Mair (1968).

ASSESSMENT TECHNIQUES IN BEHAVIOR MODIFICATION PROGRAMS

In Chapter 17, we considered some of the broad influences of social learning theory on personality assessment models. The most conspicuous effects, it will be recalled, are an increasing concern with the situational components of behavior and a growing recognition that psychological traits are not fixed, underlying, innate entities. In the present section, we shall look more specifically at the assessment needs of the behavior modification or behavior therapy programs developed through the application of social learning theory (Dickson, 1975; Evans & Nelson, 1974; Goldfried & Sprafkin, 1974; Kanfer, 1972; Kanfer & Saslow, 1969; Mischel, 1968, Ch. 8).

The principal functions to be served by assessment procedures in behavior modification programs can be subsumed under three headings. First, assessment techniques help in defining the individual's problem through a *functional analysis* of relevant behavior (Bijou & Peterson, 1971; Kanfer & Saslow, 1969). Essentially, such an analysis involves a full specification of the treatment target, such as the reduction of anxiety when speaking in public, overcoming a fear of snakes, improving self-assertive behavior, or acquiring specific social competencies. This process includes a description of the stimuli that elicit the behavior of interest, the situations in which such behavior occurs, and the nature, magnitude, and frequencies of particular responses. It is also important to investigate the conditions that tend to maintain unwanted behavior, i.e., to ascertain what are the usual consequences of the target behavior in the individual's present environment. It should be noted that the targets for behavior modification may include not only overt behavior but also cognitions and feelings.

A second requirement in behavior modification programs is the identification of effective reinforcing stimuli for use in the reconditioning process. Although such common reinforcers as food (e.g., candy), social approval, and attention from associates may often prove efficacious, a systematic survey of the relative value of available reinforcers in individual cases will enhance the probability of successful retraining. Third, there is need for assessing the behavior change that results from treatment. Such assessment should include techniques for monitoring change in the course of the program, so as to permit evaluation of treatment effectiveness and the introduction of procedural alterations at any treatment stage. It should also include terminal measures to establish the individual's attainment of satisfactory status and follow-up observations to ensure maintenance of satisfactory status.

Because of the sudden growth of behavior modification programs, the development of specialized assessment techniques to meet their needs has lagged far behind. A survey of leading clinical psychologists engaged in behavior modification revealed the use of a variety of makeshift local devices and a widespread dissatisfaction with available techniques (Kanfer, 1972). Among the procedures commonly employed are direct observations of behavior, in both naturalistic settings and specially designed standardized situations; role playing; checklists and rating scales filled out by the individual himself or by such observers as parents, teachers, classmates, and nurses or other hospital personnel; diaries of the daily occurrence of target behaviors;[5] interviewing techniques, sometimes guided by special checklists; and physiological measures, as in the assessment of anxiety or sexual arousal.

Some adaptations of early self-report inventories may be used in the preliminary screening and identification of target behaviors.[6] Standardized problem checklists, such as the Mooney Problem Check List described in Chapter 17, may also serve this function. It is characteristic of all self-report inventories employed in behavior modification programs that their responses are treated veridically and the inventories are evaluated in terms of their content validity.

A number of instruments have been prepared for use in specific research projects or treatment programs; although not themselves published, they are usually fully described and reproduced in journal articles or books. Several are fear-behavior surveys, which list items frequently found to evoke anxiety and ask the respondent to indicate the degree of disturbance aroused by each. An example is the Fear Survey Schedule (Wolpe, 1973; Wolpe & Lang, 1964), one of the few that have been published (see Appendix E). It comprises 108 items, such as the noise of vacuum cleaners, open wounds, flying insects, and taking written tests. For each item, the individual indicates how much he is disturbed by it, by marking one of five responses from "not at all" to "very much." A 40-item fear survey schedule, which yields a profile of scores in five fear categories identified through factor analysis, is described by Tasto, Hickson, and Rubin (1971).

---

[5] For an empirical study of possible sources of error in such self-records of target behavior, see Lipinski and Nelson (1974).

[6] See, e.g., Wolpe (1973, Appendix) for Bernreuter Self-Sufficiency Scale and revision of Willoughby Personality Schedule.

Another area of interest in surveys of target behaviors is that of self-assertiveness. For example, the Conflict Resolution Inventory (McFall & Lillesand, 1971; McFall & Twentyman, 1973) contains 35 items describing situations in which the individual is asked to do something unreasonable. A typical item reads:

> You are in the thick of studying for exams when a person whom you know only slightly comes into your room and says, "I'm tired of studying. Mind if I come in and take a break for a while?"

For each item the respondent indicates whether or not he would be likely to refuse the request and how comfortable he would feel about refusing or acquiescing. This technique can be used not only in the initial assessment but also in monitoring the course of assertive training.

Some instruments have been designed to aid in choosing available reinforcers for use in behavior modification programs. Among them are the Reinforcement Survey Schedule (Cautela & Kastenbaum, 1967) and the Mediator-Reinforcer Incomplete Blank (Tharp & Wetzel, 1969). The latter was constructed for use in a demonstration project on behavior modification in children. It consists of 34 incomplete sentences such as "The thing I like to do most is ______;" "I will do almost anything to get ______." In an effort to assemble a representative sample of possible reinforcing stimuli, MacPhillamy and Lewinsohn (1974, 1975) asked college students to describe "events, experiences, or activities which you find pleasant, rewarding, and fun." From the replies, they prepared the Pleasant Events Schedule, a list of 320 items for each of which the respondent indicates how often the event occurred in the past month and how enjoyable or pleasant it was. For events not recently experienced, the individual indicates how enjoyable it would have been if it had occurred.

OBSERVER REPORTS

The tests considered thus far give ample evidence of the variety of approaches that have been followed in the assessment of personality. Yet the best that can be said about most of them is that they are promising experimental instruments, suitable for research purposes, or useful adjuncts in the hands of the skilled clinician. It is apparent that for the assessment of personality variables today one cannot rely entirely on standardized tests. Other sources of information are needed to follow up or supplement the leads provided by test scores, to assess traits for which no adequate tests are available, and to obtain criterion data for developing and validating personality tests.

It is important to recognize that, especially in the domain of personality, tests cannot do the whole assessment job. Direct observations of behavior play an essential part in personality appraisal, whether in the clinic, counseling center, classroom, personnel office, or any other context calling for individual evaluations. To place such behavioral observations in the proper perspective, we must remember that all tests are themselves evaluations of small samples of behavior. To be sure, these behavior samples are obtained and evaluated under standardized conditions. But against the obvious advantages of such standardized procedures we must balance the advantages of a much more extensive sampling of behavior available through observational techniques in natural settings. To take an extreme example, if we had a detailed biography of a person extending from birth to age 30, we could probably predict his subsequent behavior more accurately than could be done with any test or battery of tests. Such a record of all the minutiae and circumstances of his life would be hard to come by; but if we had it, we could make predictions from a 30-year behavior sample rather than from the 1- or 2-hour samples provided by tests.

In all the techniques to be considered in this section, what the individual does in natural contexts over relatively long periods of time is transmitted through the medium of one or more observers. Much can be done to improve the accuracy and communicability of such observations.

NATURALISTIC OBSERVATION. Techniques for the direct observation of spontaneous behavior in natural settings have been employed most widely by child psychologists, particularly with preschool children. Although such procedures can be followed with persons of any age, the younger the individual the less likely is it that his behavior will be affected by the presence of the observer or that he has developed the social façades that complicate the interpretation of behavior. These observational techniques have also proved useful in the classroom, especially if the observer is the teacher or someone else who fits readily into the normal school setting. A major application of such assessment techniques is to be found in behavior modification programs conducted in schools, homes, child-care centers, clinics, hospitals, or any other context. Several ingenious applications of naturalistic observation have also been devised for research in social psychology (Webb, Campbell, Schwartz, & Sechrest, 1966).

Naturalistic observation covers a wide variety of procedures (Jones, Reid, & Patterson, 1975; J. S. Wiggins, 1973, pp. 295–327; Wright, 1960). They range from comprehensive, long-term techniques, as illustrated by the diary method, to more narrowly circumscribed, shorter, and more highly controlled observations, as illustrated by time sampling. The latter comprises a representative distribution of short observation periods, which has been described as "a series of far-between flashes on a behavior stream that otherwise flows in the dark" (Wright, 1960, p. 98). Depending on the nature and purpose of the observations, such periods may vary in length from less than a minute to several hours; periods of five minutes or less are the most common. The observations may be concentrated in one day or spaced over several months. They may cover all behavior occurring during the specified period; but more often they are limited to a particular kind of behavior, such as language, locomotion, interpersonal behavior, or aggression. Checklists of what to look for are a useful observational aid. Other procedural aids include observational schedules, record forms, coding systems, and mechanical recording devices (Jones, Reid, & Patterson, 1975; Lovaas, Freitag, Gold, & Kassorla, 1965; Webb et al., 1968; J. S. Wiggins, 1973). When practicable, automatic recordings can be made on tape, film, or videotape. The use of radio telemetry is also being explored for this purpose (Miklich, 1975).

It might be noted that naturalistic observations have much in common with the previously discussed situational tests. They differ principally in two respects: in naturalistic observations, no control is exerted over the stimulus situation, and—at least in most observational methods—a more extensive behavior sample is observed.

THE INTERVIEW. Mention should also be made of the time-honored source of information provided by interviewing techniques. Interviewing serves many purposes in clinical psychology, counseling, personnel psychology, and education. Psychological discussions of the methods, applications, and effectiveness of interviewing can be found in several sources.[7] In form, interviews may vary from the highly structured (representing little more than an orally administered questionnaire), through patterned or guided interviews covering certain predetermined areas, to nondirective and depth interviews in which the interviewer merely sets the stage and encourages the subject to talk as freely as possible.

Interviews provide chiefly two kinds of information. First, they afford an opportunity for direct observation of a rather limited sample of behavior manifested during the interview situation itself. For example, the individual's speech, language usage, poise, and manner in meeting a stranger can be noted. A much more important function of interviewing, however, is to elicit life-history data. What the individual has done in the past is a good indicator of what he may do in the future, especially when interpreted in the light of concomitant circumstances and of his comments regarding his actions. The interview should concern itself not only with what has happened to the individual but also with his perceptions of these events and his current evaluations of them.

---

[7] See, e.g., Bingham, Moore, and Gustad (1959), Fear (1973), Grant and Bray (1969), Kahn and Cannell (1957), Matarazzo and Wiens (1972), McCormick and Tiffin (1974, Ch. 4), Ulrich and Trumbo (1965).

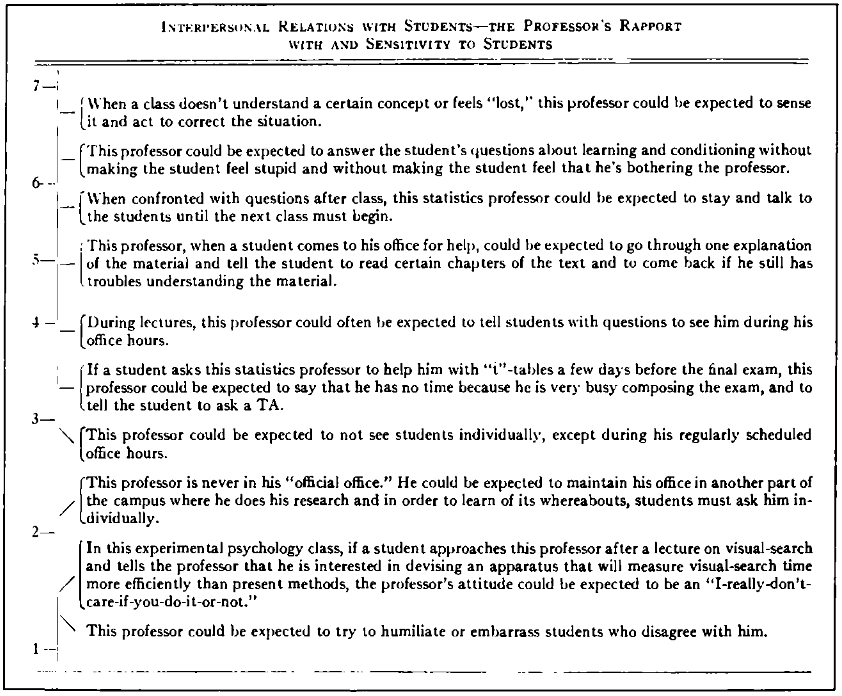
On the interviewer's part, the interview requires skill in data gathering and in data interpreting. An interview may lead to wrong decisions because important data were not elicited or because given data were inadequately or incorrectly interpreted. An important qualification of the successful interviewer is sensitivity in identifying clues in the interviewee's behavior or in facts he reports. Such clues then lead to further probing for other facts that may either support or contradict the original hypothesis.

RATINGS. Although ratings may be obtained in many contexts and for diverse purposes, the present section is concerned with the use of ratings as an evaluation of the individual by the rater on the basis of cumulative, uncontrolled observations of daily life. Such ratings differ from naturalistic observations in that the data are accumulated casually and informally; they also involve interpretation and judgment, rather than simple recording of observations. In contrast to both naturalistic observation and interviews, however, they typically cover a longer observation period and the information is obtained under more realistic conditions. Ratings are used extensively in assessing individuals in educational and industrial settings, in obtaining criterion data for test validation, and for many research purposes. In a survey of published validation data from industrial and governmental organizations, Lawshe and Balma (1966, pp. 37–38) found ratings to have been used in 68 percent of all investigations and as the sole criterion in over 60 percent.

Much can be done to improve the accuracy of ratings. A common difficulty arises from ambiguity in either trait names, scale units, or both. To meet this problem, each trait should be defined in very specific terms and the ratings should be expressed in a form that will be uniformly interpreted by all raters. Rather than using numbers or general descriptive adjectives that convey different meanings to different raters, degrees of a trait may be more clearly identified in terms of critical incidents or graded behavior samples. Smith and Kendall (1963) developed a refinement of such behaviorally anchored rating scales whereby both the qualities to be rated and the behavior samples are empirically formulated and scaled by groups of judges drawn from the same populations as the raters who will eventually use the scales. This procedure has proved effective in the development of rating scales for such diverse groups as hospital nurses, department managers in retail stores, and college professors (Campbell, Dunnette, Arvey, & Hellervik, 1973; Harari & Zedeck, 1973; Smith & Kendall, 1963; Zedeck, Imparato, Krausz, & Oleno, 1974).

As an example, we may consider a set of scales designed for student rating of psychology instructors (Harari & Zedeck, 1973). The dimensions or qualities identified through the empirical procedure were: depth of knowledge, delivery, organization, interpersonal relations with students, relevance, testing, grading, assignment and work load, and ability to motivate. One of these nine scales is reproduced in Figure 108. As in all graphic rating scales, ratings are recorded by checking any position along a continuous line. Ten scaled behavioral expectations, derived from critical incidents submitted by student groups, are shown at the appropriate position on the 7-point scale.

One of the conditions that affects the validity of ratings is the extent of the rater's *relevant acquaintance* with the person to be rated (Free-



INTERPERSONAL RELATIONS WITH STUDENTS—THE PROFESSOR'S RAPPORT
WITH AND SENSITIVITY TO STUDENTS

When a class doesn't understand a certain concept or feels "lost," this professor could be expected to sense it and act to correct the situation.

This professor could be expected to answer the student's questions about learning and conditioning without making the student feel stupid and without making the student feel that he's bothering the professor.

When confronted with questions after class, this statistics professor could be expected to stay and talk to the students until the next class must begin.

This professor, when a student comes to his office for help, could be expected to go through one explanation of the material and tell the student to read certain chapters of the text and to come back if he still has troubles understanding the material.

During lectures, this professor could often be expected to tell students with questions to see him during his office hours.

If a student asks this statistics professor to help him with "t"-tables a few days before the final exam, this professor could be expected to say that he has no time because he is very busy composing the exam, and to tell the student to ask a TA.

This professor could be expected to not see students individually, except during his regularly scheduled office hours.

This professor is never in his "official office." He could be expected to maintain his office in another part of the campus where he does his research and in order to learn of its whereabouts, students must ask him individually.

In this experimental psychology class, if a student approaches this professor after a lecture on visual-search and tells the professor that he is interested in devising an apparatus that will measure visual-search time more efficiently than present methods, the professor's attitude could be expected to be an "I-really-don't-care-if-you-do-it-or-not."

This professor could be expected to try to humiliate or embarrass students who disagree with him.

FIG. 108. A Behaviorally Anchored Graphic Rating Scale. This is one of a set of nine scales developed for student rating of faculty teaching in psychology courses.

(From Harari & Zedeck, 1973, p. 264. Copyright 1973 by the American Psychological Association. Reprinted by permission.)

berg, 1969). It is not enough to have known the person for a long time. The rater should have had an opportunity to observe him in situations in which the behavior in question could be manifested. For example, if an employee has never had an opportunity to make decisions on his job, his ability to do so cannot be evaluated by his supervisor. In many rating situations, it is desirable to include a space to be checked in lieu of a rating if the rater has had no opportunity to observe the particular trait in a given individual.

Ratings are subject to a number of constant errors. A well-known example is the *halo effect.* This is a tendency on the part of raters to be unduly influenced by a single favorable or unfavorable trait, which colors their judgment of the individual's other traits. One way to reduce halo effect is to define traits in terms of concrete behavior. In some graphic rating scales, the favorable and unfavorable ends of the line are reversed in random order for different traits. This device serves to emphasize the distinctness of each trait and forces the rater to read the anchoring statements rather than marking a similar position indiscriminately for all traits. When several persons are to be rated by the same rater, a recommended procedure is to rate all persons in a single trait before rating anyone on the next trait. Thus, all ratees are first rated in oral communication, then all are rated in decision making, and so on. This procedure tends to focus the rater's attention on each trait separately rather than on his general impression of each individual.

Another constant error is the *error of central tendency,* or the tendency to place persons in the middle of the scale and to avoid extreme positions. Still another is the *leniency error,* referring to the reluctance of many raters to assign unfavorable ratings. The former causes a bunching of ratings in the center of the scale, the latter at the upper end. Both errors reduce the effective width of the scale and make ratings less discriminative.

One way to eliminate these errors is to employ ranking or other *order-of-merit procedures.* All these procedures force discrimination among individuals and hence maximize the information yielded by the ratings. They also share the common feature of providing only relative evaluations within a group, rather than absolute evaluations against a constant, external standard. In simple ranking, someone in each group must be first and someone must be last, regardless of the quality of the group as a whole. This is also true of such techniques as paired comparisons, in which each individual is compared in turn with every other, and of forced distributions, in which the number of persons to be classified in each category is specified. All these techniques are applicable when comparisons are made within a single group; but they do not permit direct comparisons across groups evaluated by different raters.

Several other types of rating procedures have been developed to meet specific needs. Descriptions of these procedures can be found in general texts on applied or industrial psychology (see, e.g., Anastasi, 1964, Ch. 4; McCormick & Tiffin, 1974, Ch. 8). The above examples suffice to illustrate the problems encountered and some of the common solutions. In addition to improving the mechanics of rating, other steps can be taken to increase the effectiveness of ratings. Whenever feasible, having each individual evaluated by more than one rater provides a safeguard against personal bias or idiosyncrasy on the part of any one rater and also enlarges the observational domain on which the ratings are based. The rating process can also be improved by training the raters. Research in both industrial and military settings has demonstrated the effectiveness of training in increasing the validity of ratings and in reducing such judgment errors as halo effect and leniency error.

NOMINATING TECHNIQUE. A rating procedure that is especially useful in obtaining peer ratings is the nominating technique. Originally developed in sociometry (Moreno, 1953) for investigating group structure, this technique may be used within any group of persons who have been together long enough to be acquainted with one another, as in a class, factory, institution, club, or military unit. Each individual is asked to choose one or more group members with whom he would like to study, work, eat lunch, play, or carry out any other designated function. Respondents may be asked to nominate as many group members as they wish, or a specified number (such as first, second, and third choice), or only one person for each function.

When used for individual assessment, the nominations received by any one person can serve to identify potential leaders (who receive many choices) as well as isolates (who are rarely or ever mentioned). In addition, several indices can be computed for a more precise assessment of each individual. The simplest is a count of the number of times an individual was nominated for a specific function, which can be treated as his peer rating. The nominating technique can be employed with reference to any behavior of interest. For example, the respondents may be asked to name the person who has the most original ideas, who can be counted on to get the job done, or who is the best sport. They may be asked to designate not only the person who is most like the given description but also the one who is least like it. In that case, positive nominations would be weighted +1 and negative nominations −1 in totaling each person's score.

A variant of the nominating technique that is especially applicable in obtaining children's ratings of each other is the "Guess Who" technique, first used in the Character Education Inquiry together with the previously discussed situational tests. In this technique, the children are given a number of brief "word-pictures" and are instructed to write under each the name of every classmate who might fit the description. Examples include (Hartshorne, May, & Maller, 1929, p. 88):

> This is a jolly good fellow—friends with everyone, no matter who they are.
>
> This one is always picking on others and annoying them.

Using the same approach with more sophisticated psychometric procedures, Wiggins and Winder (J. S. Wiggins, 1973, pp. 358–363; Wiggins & Winder, 1961a, 1961b) developed the Peer Nomination Inventory for assessing the social adjustment of preadolescent boys. The items or descriptive statements for which individuals are nominated were themselves empirically developed from recorded interviews with some 250 8- to 12-year-old boys. The end-product was a list of 64 peer-rating items, such as "He's absent from school a lot" or "He's always losing things." The inventory yields scores in five variables: aggression, dependency, withdrawal, depression, and likeability.

However obtained, peer ratings have generally proved to be one of the most dependable of rating techniques in such diverse groups as military personnel, industrial supervisors, Peace Corps volunteers, and college students (Fiske & Cox, 1960; Hollander, 1965; Lindzey & Borgatta, 1954; Reynolds, 1966; Weitz, 1958; J. S. Wiggins, 1973, pp. 356–357). When checked against a variety of practical criteria dependent on interpersonal relations, such ratings have been found to have good predictive validity. These findings are understandable when we consider some of the features of peer ratings. First, the number of raters is large, including all group members. Second, an individual's peers are often in a particularly favorable position to observe his typical behavior. They may thus be better judges of certain interpersonal traits than teachers, supervisors, and other outside observers. Third, and probably most important, is the fact that the opinions of group members—right or wrong—influence their actions and hence partly determine the nature of the individual's subsequent interactions with the group. Other comparable groups may be expected to react toward the individual in a similar fashion. Sociometric ratings may thus be said to have content validity in the same sense as work samples.

CHECKLISTS AND Q SORTS. Any self-report instrument, such as the personality and interest inventories discussed in Chapters 17 and 18, may also be employed by an observer in describing another person. Instruments designed to assess self-concept are epecially well suited for this purpose. The Adjective Check List (ACL) has been used extensively to obtain observers' evaluations in the IPAR research program (Gough &

Heilbrun, 1965). Trained psychologists who have observed the subject closely over a 2- or 3-day assessment period record their evaluations by checking the appropriate adjectives on the list.

The Q sort has also been widely used for observer evaluations. Block (1961) describes the California Q-set (CQ-set), which was specially developed to provide a standard language for comprehensive personality evaluations by professionally trained observers. The CQ-set consists of 100 statements to be sorted into a 9-point forced distribution. The statements are sorted with regard to their "salience" for the individual—that is, their importance in specifying the unique and essential characteristics of the individual. Thus, the ipsative frame of reference typical of Q sorts is retained; the individual is not compared with outside normative standards.

The availability of such a uniform Q-set facilitates communication and assures comparability of data from different observers. The standard Q-set can also be used for a number of other research or clinical purposes. With it, evaluations can be recorded not only from personal observations but also from case histories, interviews, or projective tests. In validation research on projective techniques, for example, Q sorts have been used to obtain both criterion and test evaluations from different observers. Since the two sets of evaluations are expressed in the same terms, direct comparisons between them can be made. Another application involves the use of this technique in individual assessments. In this connection, Block (1961) provides examples of three "defining Q sorts," representing a consensus evaluation of an optimally normal individual and two psychiatric syndromes, against which a given individual's Q sort can be compared. Similar defining Q sorts may be developed for any desired category of persons. Block (1971a) has also used the Q sort in a global, developmental study of personality data from the files of two major longitudinal studies, the Oakland Growth Study and the California Guidance Study, both conducted in the San Francisco Bay Area. The age range covered by the data extends from adolescence to the mid-thirties.

BIOGRAPHICAL INVENTORIES

Both in the discussion of interviewing techniques earlier in this chapter and in the discussion of the clinician's data-gathering function in Chapter 16, reference was made to the importance of life-history data. Carefully designed biographical inventories represent an attempt to obtain some of the same types of information under uniform conditions and in situations where it would not be feasible to conduct individual interviews. In some cases, moreover, the biographical inventory may be used as a preliminary instrument to facilitate and expedite the interview.

Like the personality inventories surveyed in Chapters 17 and 18, the biographical inventory is a self-report instrument. But most of its questions generally pertain to relatively objective and readily verifiable facts. The application forms filled out by prospective employees or by students seeking admission to an educational institution are examples of biographical inventories. Typical items deal with amount and nature of education, job experiences, special skills, hobbies, and recreational activities. Frequently the individual's reaction to prior experiences is sought, as when the respondent is asked about the courses he liked best and least in school, or what he liked and disliked in his job experiences. Some biographical inventories also include items quite similar to those found in interest inventories.

For maximum predictive effectiveness, biographical inventory items should be selected and weighted by criterion keying, as in the construction of such inventories as the MMPI and the SCII discussed in Chapters 17 and 18. The resulting inventory should then be cross-validated against the same criterion in a new sample. When this procedure has been followed, biographical inventories have proved to be consistently good predictors of performance in a wide diversity of contexts. Biographical inventories have been empirically developed against such varied criteria as amount of insurance sold by life insurance salesmen, job turnover of bank clerks, productivity of research scientists, artistic creativity of high school students, and performance of naval personnel in divers training. Such inventories have proved valid as predictors of job performance in groups ranging from unskilled workers, office clerks, and service station dealers to chemists, engineers, and high-level executives.[8]

At a more basic level, Owens (1968) presents a theoretical rationale for the use of biographical inventory data to bridge the gap between psychometric and experimental approaches to psychological research. For the psychometrician, the biographical inventory can provide the antecedent information from the individual's experiential history needed for a better understanding of the correlations found among current variables, such as test scores and criterion measures. For the experimental psychologist, the biographical inventory can help to identify relatively homogeneous subgroups or types of persons, classified on the basis of their past experience. Generalizations regarding the effects of experimental variables could then be tested separately in such subgroups and possibly modified.

---

[8] Surveys of biographical inventory studies, with references, can be found in such sources as Anastasi (1964, Ch. 4), McCormick and Tiffin (1974, Ch. 4), and Owens and Henry (1966). Examples of recent investigations, not covered in those surveys, include Anastasi and Schaefer (1969), Helmreich, Bakeman, and Radloff (1973), James, Ellison, Fox, and Taylor (1974), Lee and Booth (1974), Robinson (1972), and Tanofsky, Shepps, and O'Neill (1969).

THE ASSESSMENT OF ENVIRONMENTS

APPROACHES TO ENVIRONMENTAL PSYCHOLOGY. Emerging as a distinct field of study in the mid-1960s, environmental psychology has exhibited a phenomenal growth in scope and activity. In view of the many ways in which the concept of environment enters into the science of psychology, it is not surprising that the terms "environmental psychology" and "ecological psychology" have been used to cover a wide diversity of topics (Barker, 1968; Craik, 1971, 1973; Moos, 1973; W. B. Walsh, 1973). In its more restricted sense, current environmental psychology reflects the general societal concern with such problems as conservation of animal and plant life, depletion of natural resources, and pollution of air and water. Within this area, a variety of assessment techniques are being explored, ranging from a Landscape Adjective Check List for describing natural scenery (Craik, 1971) to the Environmental Response Inventory for assessing attitudes toward such environmental concepts as pastoralism, urbanism, the need for privacy, and mechanical orientation (McKechnie, 1974).

Other aspects of environmental psychology have their roots in earlier fields of psychological research and practice. Thus, environmental psychology overlaps much of engineering psychology and systems research, which are concerned with the effects of the immediate environment on the individual's performance level and his general well-being. The relevant environmental variables pertain not only to such physical factors as lighting, temperature, air quality, noise, and equipment design and arrangement, but also to interpersonal relations and organizational climate.

A third source of interest in the assessment of environments stems from social learning theory and behavior modification techniques, with their emphasis on situational specificity (Mischel, 1968). In this approach, attention centers principally on the influence of the social environment in reinforcing specific behaviors. Information is sought on how and why an individual behaves differently in different situations. Such information should improve the prediction of the individual's behavior in particular settings. In this connection, the need for a taxonomy of situations has been recognized by several investigators, and some procedures have been proposed for working toward such a classification (N. Frederiksen, 1972).

Still another approach to environmental assessment is concerned with the cumulative effects of environment in shaping the individual's psychological development in both cognitive and noncognitive domains. Representing one aspect of the general heredity-environment problem, this field of investigation has engaged the efforts of psychologists for at least half a century. But in this area, too, there has been an upsurge of activity since the 1960s. It is characteristic of this approach that information about an individual's prior environment is used as an aid in interpreting his test performance. Moreover, as was reported in Chapter 12, the prediction of later behavior from test scores can be substantially improved if combined with knowledge about the characteristics of the environment to which the individual has been exposed during the intervening period.

OBJECTIVE INDICES OF ENVIRONMENTAL STATUS. The traditional approach to environmental assessment has relied on some global, composite index of socioeconomic level.[9] Sociologists have utilized elaborate procedures for identifying an individual's social class membership (Warner, Meeker, & Eells, 1949). Simpler and more readily applicable indices, however, have proved to be equally effective, yielding results that agree very closely with those obtained by the more laborious methods. In fact, a reasonably close approximation of socioeconomic level can be found from the occupation of the father or other principal wage earner in the family. Several rough scales have been constructed for classifying parental occupation into levels; some combine occupational information with parental educational level, as in the widely used Two-Factor Index of Social Position. First described by Hollingshead (1957), this index can be found in several sources (e.g., Bonjean et al., 1967; Stanley et al., 1972).

Other scales have attempted to give a more highly differentiated picture of home environment, including not only parental characteristics but also such data as size and nature of home; availability of such material possessions as telephones, vacuum cleaners, and refrigerators; presence of books, magazines, and newspapers in the home; and aesthetic and socio-civic involvement of the family. For these scales, data may be gathered through questionnaires, interviews, home visits, or some combination of these procedures. Examples include the American Home Scale (see Appendix E), the Home Index (Gough, 1971a, 1971b), and the Living Room Check List (Laumann & House, 1970; see also Craik, 1971, pp. 41–45).

Objective data are also frequently employed to evaluate institutions, such as industrial organizations or colleges and universities. To what extent such university characteristics as number of books in the library, faculty-student ratio, and percentage of faculty with PhD degrees actually contribute to student achievement is a matter of controversy (Astin, 1968b; Moos, 1973). A somewhat different approach describes an envi-

[9] Surveys of available indices can be found in Bonjean, Hill, and McLemore (1967), Gerberich, Greene, and Jorgensen (1962, Ch. 4), and Stanley and Hopkins (1972, pp. 411–414).

ronment in terms of the behavioral characteristics of the persons in it. The rationale underlying this procedure is that the dominant features of an environment depend on the typical characteristics of its members. The behavioral impact of a given setting is mediated by and reflected in the behavior of the persons who inhabit it. An example of this technique is provided by the Inventory of College Activities (Astin, 1968a), in which students report factual information about themselves, such as the median number of hours per week spent in different activities and the college organizations to which they belong.

OBSERVER ASSESSMENT OF SELECTED ENVIRONMENTAL VARIABLES. Another approach to environmental assessment utilizes direct observations of relevant variables by a teacher, home visitor, or other appropriate investigator. A major limitation of the traditional global indices stems from the fact that they classify environments along a single continuum of better-or-worse or higher-or-lower. Environments, however, differ in the particular behaviors they reinforce, and hence in their effects on specific individual characteristics (Bloom, 1964; Mischel, 1968). Thus the optimal environments for muscular development, school achievement, creativity, and social conformity may be quite dissimilar.

Investigators working with Bloom at the University of Chicago explored the possibility of constructing "sub-environment" scales for the prediction of selected behavior domains, namely, intelligence test performance and academic achievement (Bloom, 1964; R. Wolf, 1966). The focus was on what parents do in relation to the child, rather than on parental status and physical characteristics of the home. Examples of the variables assessed include the kinds of work habits expected of the child, social pressure and rewards for academic achievement, the opportunities provided for language development and the quality of the language models available to the child, and the nature and amount of assistance provided in overcoming academic difficulties. In some suggestive validation research with 60 fifth-grade children, the intelligence scale correlated .69 with scores on a group intelligence test, in contrast with the correlations of .20 to .40 generally found between intelligence test scores and indices of socioeconomic level. Similarly, the academic achievement scale correlated .80 with an achievement battery, in contrast with the usual correlations of the order of .50 obtained with socioeconomic level. Equally promising results were obtained in a more elaborate study with a home environment scale designed to predict performance on the SRA tests of Primary Mental Abilities in a group of Canadian school boys (Marjoribanks, 1972, 1974; Marjoribanks & Walberg, 1973; see also Harris & McArthur, 1974).

Mention should also be made of the environmental data included in the Circus assessment program for early childhood education, described in Chapter 14. Through teacher questionnaires, information is collected regarding the *context* within which the child's test scores must be interpreted (Messick, 1973a). Thus, inferences about the child's competencies from his test performance are moderated by knowledge regarding the opportunities the child has had at home and in school to learn the required skills.

PERCEIVED SOCIAL CLIMATES. Some investigators have approached the assessment of environments through the reported perceptions of the persons within each milieu. According to this theoretical orientation, the most direct measure of the social climate of a particular environment is the consensus of these individual perceptions. Every institution—be it a family, school, business organization, correctional institution, or therapeutic program—provides a social environment that tends to reinforce some types of behavior and to weaken others. The social climates of institutions, through which they influence the behavior of their members, have been described in such terms as supportive, structured, flexible, controlling, artistic, intellectual, practical.

One of the scales designed within this framework is the College Characteristics Index (Stern, 1970). Other versions of this instrument have been developed for use in high schools and evening colleges. A more general form, the Organizational Climate Index, was designed to be applicable to any organization regardless of setting. Paralleling these environmental instruments is the Activities Index, applicable to individuals. Together, the two types of instruments provide a basic taxonomy for characterizing both persons and situations in comparable terms.

The College Characteristics Index (CCI) and the Activities Index take as their point of departure Murray's (1938) theoretical model of individual needs and environmental press. The Activities Index covers 30 needs, similar to those sampled by the EPPS, the Personality Research Form, and the TAT. The CCI (and the other versions of this environmental index) measure a parallel set of "environmental press" or situational characteristics. In both types of instruments, 10 items are employed to sample each need (or press). Typical items from the CCI include:

> There would be a capacity audience for a lecture by an outstanding philosopher or theologian.
>
> The school has an excellent reputation for academic freedom.
>
> Professors usually take attendance in class.
>
> There is considerable interest in the analysis of value systems and the relativity of societies and ethics.

Data on each institution are obtained by having the CCI filled out by a sample of students. The percentages of students answering each item affirmatively or negatively are then employed in plotting the institutional profile. The index thus describes the institution as it is perceived by its students. The results for a given institution, however, are quite consistent with data obtained from other sources, such as faculty and administration. In addition to the 30 basic scores in both CCI and Activity Index, factor analyses have identified certain broader descriptive variables representing both first-order and second-order factors.

Another instrument designed for the same purpose is the College and University Environment Scales (CUES). This instrument originated in the same research project as the CCI; it differs from the CCI principally in the categories employed for describing college environments. Instead of the Murray need-press categories, CUES utilizes dimensions derived by factor analysis of institutional scores. The data for these analyses were obtained from a representative national sample of 50 colleges and universities. CUES provides scores in five factors: Practicality, Community, Awareness, Propriety, and Scholarship.

More recently, Moos (1974a, 1974b, 1975a, 1975b) has developed a series of nine Social Climate Scales applicable in the following milieus: hospital-based and community-based psychiatric treatment programs, correctional facilities, military companies, university student residences, high school classrooms, community groups, work milieus, and families. Each scale consists of 80 to 100 true-false items with which the respondent describes his perception of the given environment. Items were originally prepared to tap theoretically chosen environmental dimensions, such as press toward involvement, autonomy, or order; final terms were empirically selected on the basis of their ability to discriminate among milieus, as well as their internal consistency within subscales. Each environmental milieu is described in terms of 7 to 10 subscale scores. The scales discriminate substantially among environmental units; and they show high temporal stability and only minimal relation to personality characteristics of respondents. It is also noteworthy that the Social Climate Scales were designed for relatively small units within complex and heterogeneous institutions—for example, a classroom rather than a whole school, a treatment program rather than a whole hospital, a university student residence rather than a whole university. In this respect, these scales should yield more readily interpretable and less ambiguous data than would be obtained from a composite assessment of an entire organization.

References

ABRAHAMS, N. M., & ALF, E., JR. Pratfalls in moderator research. *Journal of Applied Psychology,* 1972, *56,* 245–251. (a)

ABRAHAMS, N. M., & ALF, E., JR. Reply to Dunnette's "Comments on Abrahams and Alf's 'Pratfalls in moderator research.'" *Journal of Applied Psychology,* 1972, *56,* 257–261. (b)

ADCOCK, C. J. *Factorial analysis for non-mathematicians.* New York: Cambridge University Press, 1954.

ADCOCK, C. J. Thematic Apperception Test. *Sixth Mental Measurements Yearbook,* 1965, 533–535.

ADKINS, D. C. *Test construction: Development and interpretation of achievement tests.* (2nd ed.) Columbus, Ohio: Merrill, 1974.

ADKINS, D. C., & BALLIF, B. L. *Animal Crackers: A Test of Motivation to Achieve. Examiner's manual.* Monterey, Calif.: CTB/McGraw-Hill, 1975.

AHLSTRÖM, K. G. Studies in spelling: I. Analysis of three different aspects of spelling ability. Report No. 20, The Institute of Education, Uppsala University (Sweden), 1964.

ALBRIGHT, L. E., SMITH, W. J., & GLENNON, J. R. A follow-up on some "invalid" tests for selecting salesmen. *Personnel Psychology,* 1959, *12,* 105–112.

ALLEN, R. M., & COLLINS, M. G. Suggestions for the adaptive administration of intelligence tests for those with cerebral palsy. *Cerebral Palsy Review,* 1955, *16,* 11–14.

ALLPORT, G. W. *Personality: A psychological interpretation.* New York: Holt, 1937.

ALLPORT, G. W., & ODBERT, H. S. Trait-names, a psycholexical study. *Psychological Monographs,* 1936, *47*(1, Whole No. 211).

AMERICAN ASSOCIATION ON MENTAL DEFICIENCY. *Adaptive Behavior Scale: Manual.* Washington, D.C.: Author, 1974.

AMERICAN INSTITUTES FOR RESEARCH. Situational tests for evaluating supervisory skill. *AIR Research Notes,* 1957, No. 14.

AMERICAN PSYCHOLOGICAL ASSOCIATION. Minutes of the annual meeting of the Council of Representatives. *American Psychologist,* 1966, *21,* 1127–1147.

AMERICAN PSYCHOLOGICAL ASSOCIATION, TASK FORCE ON EMPLOYMENT TESTING OF MINORITY GROUPS. Job testing and the disadvantaged. *American Psychologist,* 1969, *24,* 637–650.

AMES, L. B. The sequential patterning of prone progression in the human infant. *Genetic Psychology Monographs,* 1937, *19,* 409–460.

AMES, L. B., METRAUX, R. W., RODELL, J. L., & WALKER, R. N. *Rorschach responses in old age.* (rev. ed.) New York: Brunner/Mazel, 1973.

AMES, L. B., METRAUX, R. W., RODELL, J. L., & WALKER, R. N. *Child*

Rorschach responses: Developmental trends from two to ten years. (rev. ed.) New York: Brunner/Mazel, 1974.

Ames, L. B., Metraux, R. W., & Walker, R. N. *Adolescent Rorschach responses: Developmental trends from ten to sixteen years.* (2nd ed.) New York: Brunner/Mazel, 1971.

Anastasi, A. Practice and variability. *Psychological Monographs,* 1934, *45*(5, Whole No. 204).

Anastasi, A. The nature of psychological "traits." *Psychological Review,* 1948, *55,* 127–138.

Anastasi, A. The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement,* 1950, *10,* 67–78.

Anastasi, A. Age changes in adult test performance. *Psychological Reports,* 1956, *2,* 509.

Anastasi, A. *Differential psychology.* (3rd ed.) New York: Macmillan, 1958. (a)

Anastasi, A. Heredity, environment, and the question "How?" *Psychological Review,* 1958, *65,* 197–208. (b)

Anastasi, A. Psychological tests: Uses and abuses. *Teachers College Record,* 1961, *62,* 389–393.

Anastasi, A. *Fields of applied psychology.* New York: McGraw-Hill, 1964.

Anastasi, A. (Ed.) *Individual differences.* New York: Wiley, 1965.

Anastasi, A. Some implications of cultural factors for test construction. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 453–457.

Anastasi, A. Psychology, psychologists, and psychological testing. *American Psychologist,* 1967, *22,* 297–306.

Anastasi, A. *Psychological testing.* (3rd ed.) New York: Macmillan, 1968.

Anastasi, A. On the formation of psychological traits. *American Psychologist,* 1970, *25,* 899–910.

Anastasi, A. More on heritability: Addendum to the Hebb and Jensen interchange. *American Psychologist,* 1971, *26,* 1036–1037.

Anastasi, A. Technical critique. In L. A. Crooks (Ed.), *Proceedings of Invitational Conference on "An investigation of sources of bias in the prediction of job performance: A six-year study."* Princeton, N.J.: Educational Testing Service, 1972. Pp. 79–88.

Anastasi, A. Common fallacies about heredity, environment, and human behavior. *ACT Research Report* (American College Testing Program), No. 58, 1973.

Anastasi, A. Harassing a dead horse: Review of D. R. Green (Ed.), The aptitude-achievement distinction: Proceedings of the Second CTB/McGraw-Hill Conference on Issues in Educational Measurement. *Review of Education,* 1975, *1,* 356–362.

Anastasi, A., & Cordova, F. A. Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York City. *Journal of Educational Psychology,* 1953, *44,* 1–19.

Anastasi, A., & Drake, J. An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement,* 1954, *14,* 529–540.

Anastasi, A., & Schaefer, C. E. Biographical correlates of artistic and literary creativity in adolescent girls. *Journal of Applied Psychology,* 1969, 53, 267–273.

Anderson, J. E. The prediction of terminal intelligence from infant and preschool tests. *Thirty-ninth Yearbook, National Society for the Study of Education,* 1940, Part I, 385–403.

Anderson, S. B., Katz, M., & Shimberg, B. *Meeting the test.* (Rev. ed.) Englewood Cliffs, N.J.: Scholastic Book Services, 1965.

Anderson, S. B., & Messick, S. Social competency in young children. *Developmental Psychology,* 1974, *10,* 282–293.

Angoff, W. H. Scales with nonmeaningful origins and units of measurement. *Educational and Psychological Measurement,* 1962, *22,* 27–34.

Angoff, W. H. Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement,* 1964, *1,* 11–13.

Angoff, W. H. Can useful general-purpose equivalency tables be prepared for different college admissions tests? In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 251–264.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971. Ch. 15. (a)

Angoff, W. H. (Ed.) *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board, 1971. (b)

Angoff, W. H. Criterion-referencing, norm-referencing, and the SAT. *College Board Review,* 1974, 92, 3–5, 21.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement,* 1973, *10,* 95–105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test. College Entrance Examination Board, Research Report 3, 1973.

Ansbacher, H. L. The Goodenough Draw-a-Man Test and primary mental abilities. *Journal of Consulting Psychology,* 1952, *16,* 176–180.

Army Air Forces aviation psychology program, research reports. Rep. Nos. 1–19. Washington, D.C.: Government Printing Office, 1947–1948.

Arnold, G. F. A technique for measuring the mental ability of the cerebral palsied. *Psychological Service Center Journal,* 1951, *3,* 171–180.

Arthur, A. Z. Diagnostic testing and the new alternatives. *Psychological Bulletin,* 1969, *72,* 183–192.

Astin, A. W. *The college environment.* Washington, D.C.: American Council on Education, 1968. (a)

Astin, A. W. Undergraduate achievement and institutional "excellence." *Science,* 1968, *161,* 661–668. (b)

Astin, A. W., & Boruch, R. F. A "Link" system for assuring confidentiality of research data in longitudinal studies. *American Educational Research Journal,* 1970, *7,* 615–624.

ASVB counselor's manual. Universal City, Texas: Armed Forces Vocational Testing Group, 1974. 2 vols.

ATKINSON, J. W. (Ed.) *Motives in fantasy, action, and society.* Princeton, N.J.: Van Nostrand, 1958.

ATKINSON, J. W. Motivational determinants of intellective performance and cumulative achievement. In J. W. Atkinson, J. O. Raynor, et al., *Motivation and achievement.* Washington, D.C.: Winston (distr. by Halsted Press Division, Wiley), 1974. Ch. 20.

ATKINSON, J. W., FEATHER, N. T., et al. *A theory of achievement motivation.* New York: Wiley, 1966.

ATKINSON, J. W., O'MALLEY, P. M. & LENS, W. Motivation and ability: Interactive psychological determinants of intellective performance, educational achievement, and each other. In W. H. Sewell, R. M. Hauser, & D. L. Featherman (Eds.), *Schooling and achievement in American society.* New York: Academic Press, 1976. Ch. 8.

ATKINSON, R. C. Teaching children to read using a computer. *American Psychologist,* 1974, *29,* 169–178.

BAKAN, D. Clinical psychology and logic. *American Psychologist,* 1956, *11,* 655–662.

BAKARE, C. G. M. Social-class differences in the performance of Nigerian children on the Draw-a-Man test. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972. Pp. 355–363.

BAKER, F. B. Automation of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971. Ch. 8.

BALLER, W. R., CHARLES, D. C., & MILLER, E. L. Mid-life attainment of the mentally retarded: A longitudinal study. *Genetic Psychology Monographs,* 1967, *75,* 235–329.

BALMA, M. J. The concept of synthetic validity. *Personnel Psychology,* 1959, *12,* 395–396.

BALTES, P. B. Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development,* 1968, *11,* 145–171.

BANDURA, A. *Principles of behavior modification.* New York: Holt, Rinehart and Winston, 1969.

BANDURA, A., & WALTERS, R. H. *Social learning and personality development.* New York: Holt, Rinehart and Winston, 1963.

BANNISTER, D., & MAIR, J. M. M. *The evaluation of personal constructs.* New York: Academic Press, 1968.

BARKER, R. *Ecological psychology.* Stanford, Calif.: Stanford University Press, 1968.

BARRON, F. The psychology of creativity. In F. Barron et al., *New directions in psychology II.* New York: Holt, Rinehart and Winston, 1965. Pp. 3–134.

BARRON, F. *Creative person and creative process.* New York: Holt, Rinehart and Winston, 1969.

BART, W. M., & AIRASIAN, P. W. Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. *Journal of Educational Psychology,* 1974, *66,* 277–284.

BARTLETT, C. J., & EDGERTON, H. A. Stanine values for ranks for different numbers of things ranked. *Educational and Psychological Measurement,* 1966, *26,* 287–289.

BASS, A. R., & TURNER, J. N. Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology,* 1973, *57,* 101–109.

BASS, B. M. The leaderless group discussion. *Psychological Bulletin,* 1954, *51,* 465–492.

BASS, B. M. Development of a structured disguised personality test. *Journal of Applied Psychology,* 1956, *40,* 393–397.

BASS, B. M. Validity studies of a proverbs personality test. *Journal of Applied Psychology,* 1957, *41,* 158–160.

BASS, B. M. An approach to the objective assessment of successful leadership. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment.* Princeton, N.J.: Van Nostrand, 1959. Ch. 8.

BAUERNFEIND, R. H. The matter of "ipsative" scores. *Personnel and Guidance Journal,* 1962, *41,* 210–217.

BAUGHMAN, E. E. Rorschach scores as a function of examiner difference. *Journal of Projective Techniques,* 1951, *15,* 243–249.

BAUGHMAN, E. E. The role of the stimulus in Rorschach responses. *Psychological Bulletin,* 1958, *55,* 121–147.

BAUGHMAN, E. E., & DAHLSTROM, W. G. *Negro and white children: A psychological study in the rural south.* New York: Academic Press, 1968.

BAYLEY, N. On the growth of intelligence. *American Psychologist,* 1955, *10,* 805–818.

BAYLEY, N. Development of mental abilities. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology.* New York: Wiley, 1970.

BAYLEY, N., & ODEN, M. H. The maintenance of intellectual ability in gifted adults. *Journal of Gerontology,* 1955, *10,* 91–107.

BAYLEY, N., & SCHAEFER, E. S. Correlations of maternal and child behaviors with the development of mental abilities. *Monographs of the Society for Research in Child Development,* 1964, *29*(6, Serial No. 97).

BAYROFF, A. G., & FUCHS, E. F. Armed Services Vocational Aptitude Battery. U.S. Army Research Institute for the Behavioral and Social Sciences. Technical Research Report 1161, February 1970.

BECKER, W. C. The matching of behavior rating and questionnaire personality factors. *Psychological Bulletin,* 1960, *57,* 201–212.

BELL, A., & ZUBEK, J. The effect of age on the intellectual performance of mental defectives. *Journal of Gerontology,* 1960, *15,* 285–295.

BELL, F. O., HOFF, A. L., & HOYT, K. B. Answer sheets do make a difference. *Personnel Psychology,* 1964, *17,* 65–71.

BELLAK, L. *The TAT, CAT, and SAT in clinical use.* (3rd ed.) New York: Grune & Stratton, 1975.

BELLAK, L., & BELLAK, S. S. *Manual: Senior Apperception Technique.* Larchmont, N.Y.: C. P. S., 1973.

BELLAK, L., & HURVICH, M. S. A human modification of the Children's Apperception Test (CAT-H). *Journal of Projective Techniques & Personality Assessment,* 1966, *30,* 228–242.

Bemis, S. E. Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology*, 1968, *52*, 240–244.

Bender, L. A visual motor Gestalt test and its clinical use. *American Orthopsychiatric Association, Research Monographs*, 1938, No. 3.

Bennett, G. K., & Cruikshank, R. M. *A summary of clerical tests*. New York: Psychological Corporation, 1949.

Bennett, G. K., & Doppelt, J. E. *Test Orientation Procedure*. New York: Psychological Corporation, 1967.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. *Fifth edition manual for the Differential Aptitude Tests, Forms S and T*. New York: Psychological Corporation, 1974.

Bentler, P. M., Jackson, D. N., & Messick, S. Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, 1971, *76*, 186–204.

Bentler, P. M., Jackson, D. N., and Messick, S. A rose by any other name. *Psychological Bulletin*, 1972, *77*, 109–113.

Benton, A. L. *Revised Visual Retention Test: Manual*. New York: Psychological Corporation, 1974.

Berdie, R. F. Intra-individual variability and predictability. *Educational and Psychological Measurement*, 1961, *21*, 663–676.

Berdie, R. F. Intra-individual temporal variability and predictability. *Educational and Psychological Measurement*, 1969, *29*, 235–257.

Berdie, R. F., & Layton, W. L. Research on the Minnesota Counseling Inventory. *Journal of Counseling Psychology*, 1960, *7*, 218–224.

Berg, I. A. The deviation hypothesis: A broad statement of its assumptions and postulates. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967. Pp. 146–190.

Bernstein, L. The examiner as an inhibiting factor in clinical testing. *Journal of Consulting Psychology*, 1956, *20*, 287–290.

Biesheuvel, S. (Ed.) *Methods for the measurement of psychological performance*. International Biological Program (IBP) Handbook No. 10. Oxford, England: Blackwell, 1969.

Bijou, S. W., & Peterson, R. F. Functional analysis in the assessment of children. In P. McReynolds (Ed.), *Advances in psychological assessment*, Vol. 2. Palo Alto, Calif.: Science and Behavior Books, 1971. Pp. 63–78.

Binet, A., & Henri, V. La psychologie individuelle. *Année psychologique*, 1895, *2*, 411–463.

Binet, A., & Simon, Th. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année psychologique*, 1905, *11*, 191–244.

Bingham, W. V., Moore, B. V., & Gustad, J. *How to interview*. (4th ed.) New York: Harper, 1959.

Birren, J. E. Increments and decrements in the intellectual status of the aged. *Psychiatric Research Reports*, 1968, *23*, 207–214.

Blake, R. H. Industrial application of tests developed for illiterate and semiliterate people. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation*. The Hague: Mouton, 1972. Pp. 37–46.

Block, J. *The Q sort method in personality assessment and psychiatric research*. Springfield, Ill.: Charles C Thomas, 1961.

BLOCK, J. *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* New York: Appleton-Century-Crofts, 1965.

BLOCK, J. Remarks on Jackson's "review" of Block's challenge of response sets. *Educational and Psychological Measurement,* 1967, *27,* 499–502.

BLOCK, J. *Lives through time.* Berkeley, Calif.: Bancroft, 1971. (a)

BLOCK, J. On further conjectures regarding acquiescence. *Psychological Bulletin,* 1971, *76,* 205–210. (b)

BLOCK, J. The shifting definitions of acquiescence. *Psychological Bulletin,* 1972, *78,* 10–12.

BLOOM, B. S. *Stability and change in human characteristics.* New York: Wiley, 1964.

BLOOM, B. S. Learning for mastery. *Evaluation Comment* (University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs), 1968, *1,* No. 2.

BLOOM, B. S., et al. *Taxonomy of educational objectives, handbook I: Cognitive domain.* New York: McKay, 1956.

BLOOM, B. S., DAVIS, A., & HESS, R. *Compensatory education for cultural deprivation.* New York: Holt, Rinehart and Winston, 1965.

BLOOMBERG, M. (Ed.) *Creativity: Theory and research.* New Haven, Conn.: College and University Press, 1973.

BOLTON, T. L. The growth of memory in school children. *American Journal of Psychology,* 1891–1892, *4,* 362–380.

BOND, E. A. Tenth grade abilities and achievements. *Teachers College Contributions to Education,* 1940, No. 813.

BONJEAN, C. M., HILL, R. J., & MCLEMORE, S. D. *Sociological measurement: An inventory of scales and indices.* San Francisco: Chandler, 1967.

BORDIN, E. S. A theory of vocational interests as dynamic phenomena. *Educational and Psychological Measurement,* 1943, *3,* 49–65.

BORING, E. G. *A history of experimental psychology.* (Rev. ed.) New York: Appleton-Century-Crofts, 1950.

BORISLOW, B. The Edwards Personal Preference Schedule (EPPS) and fakability. *Journal of Applied Psychology,* 1958, *42,* 22–27.

BOWER, E. M. *Early identification of emotionally handicapped children in school.* (2nd ed.) Springfield, Ill.: Charles C Thomas, 1969.

BRADWAY, K. P. An experimental study of factors associated with Stanford-Binet IQ changes from the preschool to the junior high school. *Journal of Genetic Psychology,* 1945, *66,* 107–128.

BRADWAY, K. P., & ROBINSON, N. M. Significant IQ changes in twenty-five years: A follow-up. *Journal of Educational Psychology,* 1961, *52,* 74–79.

BRADWAY, K. P., THOMPSON, C. W., & CRAVENS, R. B. Preschool IQ's after twenty-five years. *Journal of Educational Psychology,* 1958, *49,* 278–281.

BRAY, D. W., & MOSES, J. L. Personnel selection. *Annual Review of Psychology,* 1972, *23,* 545–576.

BRESLOW, E. The predictive efficiency of the Law School Admission Test at the New York University School of Law. *Psychology Newsletter, New York University,* 1957, *9,* 13–22.

BRIDGEMAN, B. Effects of test score feedback on immediately subsequent test performance. *Journal of Educational Psychology,* 1974, *66,* 62–66.

BRISLIN, R. W., LONNER, W. J., & THORNDIKE, R. M. *Cross-cultural research methods.* New York: Wiley, 1974.

BROGDEN, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology,* 1946, *37,* 65–76.

BROGDEN, H. E. Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement,* 1951, *11,* 173–196.

BROWN, C. W., & GHISELLI, E. E. Percent increase in proficiency resulting from use of selective devices. *Journal of Applied Psychology,* 1953, *37,* 341–345.

BROWN, J. F. A modification of the Rosenzweig Picture-Frustration Test to study hostile interracial attitudes. *Journal of Psychology,* 1947, *24,* 247–272.

BROWN, S. R. Bibliography on Q technique and its methodology. *Perceptual and Motor Skills,* 1968, *26,* 587–613.

BUCHWALD, A. M. Values and the use of tests. *Journal of Consulting Psychology,* 1965, *29,* 49–54.

BURIK, T. E. Relative roles of the learning and motor factors involved in the digit symbol test. *Journal of Psychology,* 1950, *30,* 33–42.

BURKE, H. R. Raven's Progressive Matrices: A review and critical evaluation. *Journal of Genetic Psychology,* 1958, *93,* 199–228.

BURNHAM, P. S. Prediction and performance. In *From high school to college: Readings for counselors.* New York: College Entrance Examination Board, 1965. Pp. 65–71.

BURNHAM, P. S., & CRAWFORD, A. B. Law school prediction at mid-century. *Journal of Legal Education,* 1957, *10,* 201–207.

BURNS, R. B. Age and mental ability: Re-testing with thirty-three years' interval. *British Journal of Educational Psychology,* 1966, *36,* 116.

BUROS, O. K. (Ed.) *Reading tests and reviews.* Highland Park, N.J.: Gryphon Press, 1968.

BUROS, O. K. (Ed.) *Personality tests and reviews.* Highland Park, N.J.: Gryphon Press, 1970.

BUROS, O. K. (Ed.) *Seventh Mental Measurements Yearbook.* Highland Park, N.J.: Gryphon Press, 1972.

BUROS, O. K. (Ed.) *Tests in print* II. Highland Park, N.J.: Gryphon Press, 1974.

BURT, C. *The factors of the mind: An introduction to factor-analysis in psychology.* New York: Macmillan, 1941.

BURT, C. Mental abilities and mental factors. *British Journal of Educational Psychology,* 1944, *14,* 85–89.

BURT, C. The structure of the mind; a review of the results of factor analysis. *British Journal of Educational Psychology,* 1949, *19,* 110–111; 176–199.

BURTON, R. V. Generality of honesty reconsidered. *Psychological Review,* 1963, *70,* 481–499.

Burtt, H. E. *Legal psychology.* Englewood Cliffs, N.J.: Prentice-Hall, 1931.

Buss, A. R. An extension of developmental models that separate ontogenetic changes and cohort differences. *Psychological Bulletin,* 1973, *80,* 466–479.

Butcher, J. N. (Ed.) *MMPI: Research developments and clinical applications.* New York: McGraw-Hill, 1969.

Butcher, J. N. (Ed.) *Objective personality assessment: Changing perspectives.* New York: Academic Press, 1972.

Butcher, J., Ball, B., & Ray, E. Effects of socioeconomic level on MMPI differences in Negro-white college students. *Journal of Counseling Psychology,* 1964, *11,* 83–87.

Caldwell, M. B., & Smith, T. A. Intellectual structure of southern Negro children. *Psychological Reports,* 1968, *23,* 63–71.

Caldwell, O. W., & Courtis, S. A. *Then and now in education,* 1845–1923. Yonkers, N.Y.: World Book Co., 1923.

Campbell, D. P. A cross-sectional and longitudinal study of scholastic abilities over twenty-five years. *Journal of Counseling Psychology,* 1965, *12,* 55–61.

Campbell, D. P. The Minnesota Vocational Interest Inventory. *Personnel and Guidance Journal,* 1966, *44,* 854–858.

Campbell, D. P. *Handbook for the Strong Vocational Interest Blank.* Stanford, Calif.: Stanford University Press, 1971.

Campbell, D. P. *Manual for the Strong-Campbell Interest Inventory.* Stanford, Calif.: Stanford University Press, 1974.

Campbell, D. T. The indirect assessment of social attitudes. *Psychological Bulletin,* 1950, *47,* 15–38.

Campbell, D. T. A typology of tests, projective and otherwise. *Journal of Consulting Psychology,* 1957, *21,* 207–210.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist,* 1960, *15,* 546–553.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology,* 1973, *57,* 15–22.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. *An investigation of sources of bias in the prediction of job performance: A six-year study.* Princeton, N.J.: Educational Testing Service, 1973.

Canfield, A. A. The "sten" scale—A modified C-scale. *Educational and Psychological Measurement,* 1951, *11,* 295–297.

Carnes, G. D. Vocational interest characteristic of abnormal personalities. *Journal of Counseling Psychology,* 1964, *11,* 272–279.

Carroll, H. A. What do the Meier-Seashore and the McAdory Art Tests measure? *Journal of Educational Research,* 1933, *26,* 661–665.

Carroll, J. B. A model of school learning. *Teachers College Record,* 1963, *64,* 723–733.

Carroll, J. B. Factors of verbal achievement. In A. Anastasi (Ed.), *Testing* *problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 406–413.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Educational Horizons,* 1970, *48,* 71–80.

Carter, L. F. Evaluating the performance of individuals as members of small groups. *Personnel Psychology,* 1954, *7,* 477–484.

Casebook on ethical standards of psychologists. Washington, D.C.: American Psychological Association, 1967.

Cattell, J. McK. Mental tests and measurements. *Mind,* 1890, *15,* 373–380.

Cattell, R. B. *Description and measurement of personality.* Yonkers, N.Y.: World Book Co., 1946.

Cattell, R. B. *Personality and motivation structure and measurement.* Yonkers, N.Y.: World Book Co., 1957.

Cattell, R. B. *Handbook of modern personality theory.* Chicago: Aldine, 1971.

Cattell, R. B. *Personality and mood by questionnaire.* San Francisco: Jossey-Bass, 1973.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. *Handbook for the Sixteen Personality Factor Questionnaire.* Champaign, Ill.: Institute for Personality and Ability Testing, 1970.

Cattell, R. B., & Luborsky, L. B. Personality factors in response to humor. *Journal of Abnormal and Social Psychology,* 1947, *42,* 402–421.

Cattell, R. B., & Warburton, F. *Principles of objective personality measurement, and a compendium of objective tests.* Urbana, Ill.: University of Illinois Press, 1967.

Cautela, J. R., & Kastenbaum, R. A reinforcement survey schedule for use in therapy, training, and research. *Psychological Reports,* 1967, *20,* 1115–1130.

Chapman, L. J. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior,* 1967, *6,* 151–155.

Chapman, L. J., & Chapman, J. P. Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology,* 1967, *72,* 193–204.

Chapman, L. J., & Chapman, J. P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology,* 1969, *74,* 271–280.

Charles, D. C. Ability and accomplishment of persons earlier judged mentally deficient. *Genetic Psychology Monographs,* 1953, *47,* 3–71.

Charles, D. C., & James, S. T. Stability of average intelligence. *Journal of Genetic Psychology,* 1964, *105,* 105–111.

Christal, R. E. Factor analytic study of visual memory. *Psychological Monographs,* 1958, *72*(13, Whole No. 466).

Chun, Ki-Taek, Cobb, S., & French, J. R. P., Jr. *Measures for psychological assessment.* Ann Arbor, Mich.: Institute for Social Research, The University of Michigan, 1976.

Circus: Comprehensive assessment in nursery school and kindergarten. (Proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973) Princeton, N.J.: Educational Testing Service, 1973.

CLARK, J. H. The interpretation of the MMPI profiles of college students; mean scores for male and female groups. *Journal of Social Psychology*, 1954, *40*, 319–321.

CLARK, K. E. *The vocational interests of nonprofessional men.* Minneapolis: University of Minnesota Press, 1961.

CLARK, K. E., & CAMPBELL, D. P. *Manual for the Minnesota Vocational Interest Inventory.* New York: Psychological Corporation, 1965.

CLEARY, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115–124.

CLEARY, T. A., & HILTON, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, *28*, 61–75.

CLEARY, T. A., HUMPHREYS, L. G., KENDRICK, S. A., & WESMAN, A. Educational uses of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15–41. (See also comments on pp. 88–96.)

CLEARY, T. A., LINN, R. L., & ROCK, D. A. An exploratory study of programmed tests. *Educational and Psychological Measurement*, 1968, *28*, 347–349.

CLEMANS, W. V. An index of item-criterion relationship. *Educational and Psychological Measurement*, 1958, *18*, 167–172.

CLEMANS, W. V. Test administration. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C. American Council on Education, 1971. Ch. 7.

COHEN, E. Examiner differences with individual intelligence tests. *Perceptual and Motor Skills*, 1965, *20*, 1324.

COHEN, J. A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology*, 1957, *21*, 451–457. (a)

COHEN, J. The factorial structure of the WAIS between early adulthood and old age. *Journal of Consulting Psychology*, 1957, *21*, 283–290. (b)

COHEN, R. A. Conceptual styles, culture conflict, and nonverbal tests. *American Anthropologist*, 1969, *71*, 828–856.

COLE, N. S. Bias in selection. *American College Testing Program, ACT Research Report* No. 51, May 1972.

COLE, S., & HUNTER, M. Pattern analysis of WISC scores achieved by culturally disadvantaged children. *Psychological Reports*, 1971, *29*, 191–194.

COLEMAN, W., & CURETON, E. E. Intelligence and achievement: The "jangle fallacy" again. *Educational and Psychological Measurement*, 1954, *14*, 347–351.

COLLEGE ENTRANCE EXAMINATION BOARD. *Effects of coaching on Scholastic Aptitude Test scores.* New York: Author, 1968.

COLLEGE ENTRANCE EXAMINATION BOARD. *About the achievement tests*, 1974–75. New York: Author, 1974. (a)

COLLEGE ENTRANCE EXAMINATION BOARD. *About the SAT, 1974–75.* New York: Author, 1974. (b)

COLLINS, B. E. Four components of the Rotter Internal-External Scale: Belief in a difficult world, a just world, a predictable world, and a politically responsive world. *Journal of Personality and Social Psychology,* 1974, *29,* 381–391.

Comrey, A. L., Backer, T. E., & Glaser, E. M. *A sourcebook for mental health measures.* Los Angeles: Human Interaction Research Institute, 1973.

Cooley, W. W., & Glaser, R. The computer and individualized instruction. *Science,* 1969, *166,* 574–582.

Corah, M. L., et al. Social desirability as a variable in the Edwards Personal Preference Schedule. *Journal of Consulting Psychology,* 1958, *22,* 70–72.

Costello, J., & Ali, F. Reliability and validity of Peabody Picture Vocabulary Test scores of disadvantaged preschool children. *Psychological Reports,* 1971, *28,* 755–760.

Couch, A., & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology,* 1960, *60,* 151–174.

Cozby, P. C. Self-disclosure: A literature review. *Psychological Bulletin,* 1973, *79,* 73–91.

Craik, K. H. The assessment of places. In P. McReynolds (Ed.), *Advances in psychological assessment.* Palo Alto, Calif.: Science and Behavior Books, 1971. Vol. 2, Ch. 3.

Craik, K. H. Environmental psychology. *Annual Review of Psychology,* 1973, *24,* 403–421.

Creager, J. A. Predicting doctorate attainment with GRE and other variables. Technical Report No. 25, Office of Scientific Personnel, National Academy of Sciences–National Research Council, Washington, D.C., 1965.

Cronbach, L. J. Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin,* 1949, *46,* 393–429.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951, *16,* 297–334.

Cronbach, L. J., & Drenth, P. J. D. (Eds.) *Mental tests and cultural adaptation.* The Hague: Mouton, 1972.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* (2nd ed.) Urbana, Ill.: University of Illinois Press, 1965.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, 52, 281–302.

Cronbach, L. J., & Warrington, W. G. Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika,* 1951, *16,* 167–188.

Cronbach, L. J., & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika,* 1952, *17,* 124–147.

Crowne, D. P., & Marlowe, D. *The approval motive: Studies in evaluative dependence.* New York: Wiley, 1964.

Cundick, B. P. Measures of intelligence on Southwest Indian students. *Journal of Social Psychology,* 1970, *81,* 151–156.

Cureton, E. E. Validity, reliability, and baloney. *Educational and Psychological Measurement,* 1950, *10,* 94–96.

Cureton, E. E. Recipe for a cookbook. *Psychological Bulletin,* 1957, *54,* 494–497. (a)

Cureton, E. E. The upper and lower twenty-seven percent rule. *Psychometrika,* 1957, *22,* 293–296. (b)

Cureton, E. E. Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement,* 1965, *25,* 327–346.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook.* Vol. I, *Clinical interpretation.* Minneapolis: University of Minnesota Press, 1972.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook.* Vol. II, *Research developments and applications.* Minneapolis: University of Minnesota Press, 1975.

Dailey, C. A. The life history as a criterion of assessment. *Journal of Counseling Psychology,* 1960, *7,* 20–23.

Damon, A. Discrepancies between findings of longitudinal and cross-sectional studies in adult life: Physique and physiology. *Human Development,* 1965, *8,* 16–22.

D'Andrade, R. G. Trait psychology and componential analysis. *American Anthropologist,* 1965, *67,* 215–228.

Darley, J. G., & Hagenah, T. *Vocational interest measurement: Theory and practice.* Minneapolis: University of Minnesota Press, 1955.

Darlington, R. B., & Stauffer, G. F. A method for choosing a cutting point on a test. *Journal of Applied Psychology,* 1966, *50,* 229–231.

Das, R. S. Analysis of the components of reasoning in nonverbal tests and the structure of reasoning in a bilingual population. *Archiv für die gesamte Psychologie,* 1963, *115*(3), 217–229.

Davids, A., & Pildner, H., Jr. Comparison of direct and projective methods of personality assessment under different conditions of motivation. *Psychological Monographs,* 1958, 72(11, Whole No. 464).

Davis, P. C. A factor analysis of the Wechsler-Bellevue Scale. *Educational and Psychological Measurement,* 1956, *16,* 127–146.

Davis, W. E. Effect of prior failure on subjects' WAIS Arithmetic subtest scores. *Journal of Clinical Psychology,* 1969, *25,* 72–73. (a)

Davis, W. E. Examiner differences, prior failure, and subjects' arithmetic scores. *Journal of Clinical Psychology,* 1969, *25,* 178–180. (b)

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin,* 1974, *81,* 95–106.

Dearborn, W. F., & Rothney, J. *Predicting the child's development.* Cambridge, Mass.: Sci-Art Pub., 1941.

Demming, J. A., & Pressey, S. L. Tests "indigenous" to the adult and older years. *Journal of Counseling Psychology,* 1957, *4,* 144–148.

Dennis, W. Goodenough scores, art experience, and modernization. *Journal of Social Psychology,* 1966, *68,* 211–228.

Denny, J. P. Effects of anxiety and intelligence on concept formation. *Journal of Experimental Psychology,* 1966, *72,* 596–602.

Deutsch, M., Fishman, J. A., Kogan, L., North, R., & Whiteman, M. Guidelines for testing minority group children. *Journal of Social Issues,* 1964, *20,* 127–145.

DeWitt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement. Psychometric Methods Program, Department of Psychology, University of Minnesota, Research Report 74–1, January 1974.

Dicken, C. F. Simulated patterns on the Edwards Personal Preference Schedule. *Journal of Applied Psychology,* 1959, *43,* 372–378.

Dickson, C. R. Role of assessment in behavior therapy. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975. Ch. 9.

Doll, E. A. (Ed.) *The Oseretsky Tests of Motor Proficiency.* Minneapolis: American Guidance Service, 1946.

Doll, E. A. *The measurement of social competence.* Minneapolis: American Guidance Service, 1953.

Doll, E. A. *Vineland Social Maturity Scale: Manual of directions.* (Rev. ed.) Minneapolis: American Guidance Service, 1965.

Donofrio, A. F. Clinical value of infant testing. *Perceptual and Motor Skills,* 1965, *21,* 571–574.

Doppelt, J. E. Estimating the Full Scale score of the Wechsler Adult Intelligence Scale from scores on four subtests. *Journal of Consulting Psychology,* 1956, *20,* 63–66.

Doppelt, J. E., & Wallace, W. L. Standardization of the Wechsler Adult Intelligence Scale for older persons. *Journal of Abnormal and Social Psychology,* 1955, *51,* 312–330.

Dorcus, R. M., & Jones, M. H. *Handbook of employee selection.* New York: McGraw-Hill, 1950.

Drake, L. E., & Oetting, E. R. *An MMPI codebook for counselors.* Minneapolis: University of Minnesota Press, 1959.

Drasgow, J., & Carkhuff, R. R. Kuder neuropsychiatric keys before and after psychotherapy. *Journal of Counseling Psychology,* 1964, *11,* 67–71.

Dreger, R. M. General temperament and personality factors related to intellectual performances. *Journal of Genetic Psychology,* 1968, *113,* 275–293.

Droege, R. C. Effects of practice on aptitude scores. *Journal of Applied Psychology,* 1966, *50,* 306–310.

Droege, R. C. GATB longitudinal validation study. *Journal of Counseling Psychology,* 1968, *15,* 41–47.

DuBois, P. H. (Ed.) *The classification program.* (AAF Aviation Psychology Program, Research Reports. Rep. No. 2) Washington, D.C.: Government Printing Office, 1947.

DuBois, P. H. A test-dominated society: China 1115 B.C.–1905 A.D. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 29–36.

DuBois, P. H. *A history of psychological testing.* Boston: Allyn and Bacon, 1970.

DuBois, P. H. College Board Scholastic Aptitude Test. *Seventh Mental Measurements Yearbook,* 1972, 646–648.

Dunn, J. A. Inter- and intra-rater reliability of the new Harris-Goodenough Draw-A-Man Test. *Perceptual and Motor Skills,* 1967, *24,* 269–270.

Dunnette, M. D. Use of the sugar pill by industrial psychologists. *American Psychologist,* 1957, *12,* 223–225.

Dunnette, M. D. A note on *the* criterion. *Journal of Applied Psychology,* 1963, *47,* 251–254.

Dunnette, M. D. Comments on Abrahams and Alf's "Pratfalls in moderator research." *Journal of Applied Psychology,* 1972, *56,* 252–256.

Dunteman, G. H., & Bailey, J. P., Jr. A canonical correlational analysis of the Strong Vocational Interest Blank and the Minnesota Multiphasic Personality Inventory for a female college population. *Educational and Psychological Measurement,* 1967, *27,* 631–642.

Durnall, E. J., Jr. Falsification of interest patterns of the Kuder Preference Record. *Journal of Educational Psychology,* 1954, *45,* 240–243.

Dvorak, B. J. The General Aptitude Test Battery. *Personnel and Guidance Journal,* 1956, *35,* 145–154.

Dyer, H. S. Recycling the problems in testing. *Proceedings of the 1972 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1973. Pp. 85–95.

Ebbinghaus, H. Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. *Zeitschrift für angewandte Psychologie,* 1897, *13,* 401–459.

Ebel, R. L. Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement,* 1956, *16,* 269–282.

Ebel, R. L. Must all tests be valid? *American Psychologist,* 1961, *16,* 640–647.

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement,* 1962, *22,* 15–25.

Ebel, R. L. *Measuring educational achievement.* Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Ebel, R. L. The relation of item discrimination to test reliability. *Journal of Educational Measurement,* 1967, *4,* 125–128.

Ebel, R. L. *Essentials of educational measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1972. (a)

Ebel, R. L. Some limitations of criterion-referenced measurement. In G. H. Bracht, K. D. Hopkins, & J. C. Stanley (Eds.), *Perspectives in educational and psychological measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1972. Pp. 144–149. (b)

Ebel, R. L., & Damrin, D. E. Tests and examinations. *Encyclopedia of Educational Research,* 3rd ed., 1960, 1502–1517.

Edgerton, H. A. A table for computing the phi coefficient. *Journal of Applied Psychology,* 1960, *44,* 141–145.

Edwards, A. J. Using vocabulary as a measure of general ability. *Personnel and Guidance Journal,* 1963, *42,* 153–154.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957. (a)

Edwards, A. L. *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts, 1957. (b)

Edwards, A. L. Relationship between probability of endorsement and social desirability scale value for a set of 2,824 personality statements. *Journal of Applied Psychology,* 1966, *50,* 238–239.

Edwards, A. L. *The measurement of personality traits by scales and inventories.* New York: Holt, Rinehart and Winston, 1970.

EELLS, K., et al. *Intelligence and cultural differences.* Chicago: University of Chicago Press, 1951.

EISDORFER, C. The WAIS performance of the aged: A retest evaluation. *Journal of Gerontology,* 1963, *18,* 169–172.

EISDORFER, C., & COHEN, L. D. The generality of the WAIS standardization for the aged: A regional comparison. *Journal of Abnormal and Social Psychology,* 1961, *62,* 520–527.

ENDLER, N. S., & HUNT, J. McV. Sources of behavioral variance as measured by the S-R Inventory of Anxiousness. *Psychological Bulletin,* 1966, *65,* 336–346.

ENDLER, N. S., & HUNT, J. McV. S-R inventories of hostility and comparisons of the proportions of variance from persons, responses, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology,* 1968, *9,* 309–315.

ENDLER, N. S., & HUNT, J. McV. Generalizability of contributions from sources of variance in the S-R inventories of anxiousness. *Journal of Personality,* 1969, *37,* 1–24.

ENDLER, N. S., HUNT, J. McV., & ROSENSTEIN, A. J. An S-R inventory of anxiousness. *Psychological Monographs,* 1962, *76*(17, Whole No. 536).

ENGELHART, M. D. A comparison of several item discrimination indices. *Journal of Educational Measurement,* 1965, *2,* 69–76.

ENTWISLE, D. R. To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin,* 1972, *77,* 377–391.

EPSTEIN, S. Some theoretical considerations on the nature of ambiguity and the use of stimulus dimensions in projective techniques. *Journal of Consulting Psychology,* 1966, *30,* 183–192.

EQUAL EMPLOYMENT OPPORTUNITY COMMISSION (EEOC). Guidelines on employee selection procedures. *Federal Register,* 1970, *35*(149), 12333–12336.

ERIKSON, E. H. *Childhood and society.* New York: Norton, 1950.

ESCALONA, S. K. The use of infant tests for predictive purposes. *Bulletin of the Menninger Clinic,* 1950, *14,* 117–128.

ESQUIROL, J. E. D. *Des maladies mentales considérées sous les rapports médical, hygiénique, et médico-légal.* Paris: Baillière, 1838. 2 vols.

ESTES, B. W. Influence of socioeconomic status on Wechsler Intelligence Scale for Children: An exploratory study. *Journal of Consulting Psychology,* 1953, *17,* 58–62.

ESTES, B. W. Influence of socioeconomic status on Wechsler Intelligence Scale for Children: Addendum. *Journal of Consulting Psychology,* 1955, *19,* 225–226.

Ethical principles in the conduct of research with human participants. Washington, D.C.: American Psychological Association, 1973.

EVANS, I. M., & NELSON, R. O. A curriculum for the teaching of behavior assessment. *American Psychologist,* 1974, *29,* 598–606.

EXNER, J. E., JR. Variations in WISC performances as influenced by differences in pretest rapport. *Journal of General Psychology,* 1966, *74,* 299–306.

FAN, C. T. *Item analysis table.* Princeton, N.J.: Educational Testing Service, 1952.

FAN, C. T. Note on construction of an item analysis table for the high-low-27-per-cent group method. *Psychometrika*, 1954, *19*, 231–237.

FARNSWORTH, P. R. Shifts in the values of opinion items. *Journal of Psychology*, 1943, *16*, 125–128.

FEAR, R. A. *The evaluation interview.* (2nd ed.) New York: McGraw-Hill, 1973.

FELDHUSEN, J. F., & KLAUSMEIER, H. J. Anxiety, intelligence, and achievement in children of low, average, and high intelligence. *Child Development*, 1962, *33*, 403–409.

FELDMAN, D. H., & BRATTON, J. C. Relativity and giftedness: Implications for equality of educational opportunity. *Exceptional Children*, 1972, *38*, 491–492.

FERGUSON, G. A. On learning and human ability. *Canadian Journal of Psychology*, 1954, *8*, 95–112.

FERGUSON, G. A. On transfer and the abilities of man. *Canadian Journal of Psychology*, 1956, *10*, 121–131.

FERGUSON, R. L., & NOVICK, M. R. Implementation of a Bayesian system for decision analysis in a program of individually prescribed instruction. *ACT Research Report* (American College Testing Program). No. 60, 1973.

FIELD, J. G. Two types of tables for use with Wechsler's Intelligence Scales. *Journal of Clinical Psychology*, 1960, *16*, 3–6.

FIGURELLI, J. C., & KELLER, H. R. The effects of training and socioeconomic class upon the acquisition of conservation concepts. *Child Development*, 1972, *43*, 293–298.

FINCHER, C. Personnel testing and public policy. *American Psychologist*, 1973, *28*, 489–497.

FINDLEY, W. G. Rationale for the evaluation of item discrimination statistics. *Educational and Psychological Measurement*, 1956, *16*, 175–180.

FINK, A. M., & BUTCHER, J. N. Reducing objections to personality inventories with special instructions. *Educational and Psychological Measurement*, 1972, *32*, 631–639.

FISHBEIN, M. (Ed.) *Readings in attitude theory and measurement.* New York: Wiley, 1967.

FISHER, J. The twisted pear and the prediction of behavior. *Journal of Consulting Psychology*, 1959, *23*, 400–405.

FISHMAN, J. A., & LORGE, I. The influence of judges' characteristics on item judgments and on Thurstone scaling via the method of ranks (utilization of judges with varying national, religious, and experiential backgrounds). *Journal of Social Psychology*, 1959, *49*, 187–205.

FISKE, D. W. Can a personality construct be validated empirically? *Psychological Bulletin*, 1973, *80*, 89–92.

FISKE, D. W., & BAUGHMAN, E. E. Relationships between Rorschach scoring categories and the total number of responses. *Journal of Abnormal and Social Psychology*, 1953, *48*, 25–32.

FISKE, D. W., & COX, J. A., JR. The consistency of ratings by peers. *Journal of Applied Psychology*, 1960, *44*, 11–17.

FITZGIBBON, T. J. *The use of standardized instruments with urban and minority-group pupils.* New York: Harcourt Brace Jovanovich, 1972.

FJELD, S. P., & LANDFIELD, A. W. Personal construct consistency. *Psychological Reports,* 1961, *8,* 127–129.

FLANAGAN, J. C. Scientific development of the use of human resources: Progress in the Army Air Forces. *Science,* 1947, *105,* 57–60.

FLANAGAN, J. C. Critical requirements: A new approach to employee evaluation. *Personnel Psychology,* 1949, *2,* 419–425.

FLANAGAN, J. C. The critical incident technique. *Psychological Bulletin,* 1954, *51,* 327–358. (a)

FLANAGAN, J. C. Some considerations in the development of situation tests. *Personnel Psychology,* 1954, *7,* 461–464. (b)

FLANAGAN, J. C. Symposium: Standard scores for aptitude and achievement tests: Discussion. *Educational and Psychological Measurement,* 1962, *22,* 35–39.

FLANAGAN, J. C. The PLAN system for individualizing education. *NCME Measurement in Education,* 1971, *2*(2), 1–8.

FLANAGAN, J. C., et al. *The American high school student.* Pittsburgh: Project TALENT Office, University of Pittsburgh, 1964.

FLANAGAN, J. C., SHANNER, W. M., BRUDNER, H. J., & MARKER, R. W. An individualized instructional system: PLAN*. In H. Talmage (Ed.), *Systems of individualized education.* Berkeley, Calif.: McCutchan, 1975. Ch. 5.

FLAUGHER, R. L. Some points of confusion in discussing the testing of black students. In L. P. Miller (Ed.), *The testing of black students: A symposium.* Englewood Cliffs, N.J.: Prentice-Hall, 1974. Pp. 11–16.

FLAVELL, J. H. *The developmental psychology of Jean Piaget.* Princeton, N.J.: Van Nostrand, 1963.

FLEISHMAN, E. A. Dimensional analysis of psychomotor abilities. *Journal of Experimental Psychology,* 1954, *48,* 437–454.

FLEISHMAN, E. A. Dimensional analysis of movement reactions. *Journal of Experimental Psychology,* 1958, *55,* 438–453.

FLEISHMAN, E. A. The description and prediction of perceptual-motor skill learning. In R. Glaser (Ed.), *Training, research, and education.* Pittsburgh: University of Pittsburgh Press, 1962.

FLEISHMAN, E. A. *The structure and measurement of physical fitness.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

FLEISHMAN, E. A. On the relation between abilities, learning, and human performance. *American Psychologist,* 1972, *27,* 1018–1032.

FLEISHMAN, E. A., & ELLISON, G. D. A factor analysis of fine manipulative tests. *Journal of Applied Psychology,* 1962, *46,* 96–105.

FLEISHMAN, E. A., & FRUCHTER, B. Factor structure and predictability of successive stages of learning Morse code. *Journal of Applied Psychology,* 1960, *44,* 96–101.

FLEISHMAN, E. A., & HEMPEL, W. E., JR. Factorial analyses of complex psychomotor performance and related skills. *Journal of Applied Psychology,* 1956, *40,* 96–104.

FORCE, R. C. Development of a covert test for the detection of alcoholism by a keying of the Kuder Preference Record, *Quarterly Journal of Studies on Alcohol,* 1958, *19,* 72–78.

Forer, B. The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology,* 1949, *44,* 118–123.

Fossum, J. A. An application of techniques to shorten tests and increase validity. *Journal of Applied Psychology,* 1973, *57,* 90–92.

Franks, C. M. (Ed.) *Behavior therapy: Appraisal and status.* New York: McGraw-Hill, 1969.

Frederiksen, C. H. Abilities, transfer, and information retrieval in verbal learning. *Multivariate Behavioral Research Monographs,* 1969. No. 69–2.

Frederiksen, N. Response set scores as predictors of performance. *Personnel Psychology,* 1965, *18,* 225–244.

Frederiksen, N. In-basket tests and factors in administrative performance. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 208–221. (a)

Frederiksen, N. Validation of a simulation technique. *Organizational Behavior & Human Performance,* 1966, *1,* 87–109. (b)

Frederiksen, N. Toward a taxonomy of situations. *American Psychologist,* 1972, *27,* 114–123.

Frederiksen, N., & Gilbert, A. C. F. Replication of a study of differential predictability. *Educational and Psychological Measurement,* 1960, *20,* 759–767.

Frederiksen, N., & Melville, S. D. Differential predictability in the use of test scores. *Educational and Psychological Measurement,* 1954, *14,* 647–656.

Frederiksen, N., Saunders, D. R., & Wand, B. The In-Basket Test. *Psychological Monographs,* 1957, *71*(9, Whole No. 438).

Freeberg, N. E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology,* 1969, *53,* 518–524.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs,* 1951, No. 5.

French, J. W. Effect of anxiety on verbal and mathematical examination scores. *Educational and Psychological Measurement,* 1962, *22,* 553–564.

French, J. W. The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement,* 1965, *25,* 9–28.

French, J. W. The logic of and assumptions underlying differential testing. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 321–330.

French, J. W., Ekstrom, R. B., & Price, L. A. *Kit of reference tests for cognitive factors.* (Rev. ed.) Princeton, N.J.: Educational Testing Service, 1963.

Fromm, E., Hartman, L. D., & Marschak, M. Children's intelligence tests as a measure of dynamic personality functioning. *American Journal of Orthopsychiatry,* 1957, *27,* 134–144.

Fromme, A. On the use of certain qualitative methods of attitude research. *Journal of Social Psychology,* 1941, *13,* 425–459.

Fryer, D. *Measurement of interests.* New York: Holt, 1931.

FULKERSON, S. C. Individual differences in response validity. *Journal of Clinical Psychology,* 1959, *15,* 169–173.

GAGE, N. L., & CRONBACH, L. J. Conceptual and methodological problems in interpersonal perception. *Psychological Review,* 1955, *62,* 411–422.

GAGNÉ, R. *The conditions of learning.* New York: Holt, Rinehart and Winston, 1965.

GALTON, F. Psychometric experiments. *Brain,* 1879, *2,* 149–162.

GALTON, F. *Inquiries into human faculty and its development.* London: Macmillan, 1883.

GAMBLE, K. R. The Holtzman Inkblot Technique: A review. *Psychological Bulletin,* 1972, *77,* 172–194.

GARDNER, J. W. *Excellence.* New York: Harper, 1961.

GAUDRY, E., & SPIELBERGER, C. D. *Anxiety and educational achievement.* New York: Wiley, 1974.

GERBERICH, J. R., GREENE, H. A., & JORGENSEN, A. N. *Measurement and evaluation in the modern school.* New York: McKay, 1962.

GESELL, A., et al. *The first five years of life.* New York: Harper, 1940.

GESELL, A., & AMATRUDA, C. S. *Developmental diagnosis.* (2nd ed.) New York: Hoeber-Harper, 1947.

GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology,* 1956, *40,* 374–377.

GHISELLI, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement,* 1960, *20,* 675–684. (a)

GHISELLI, E. E. The prediction of predictability. *Educational and Psychological Measurement,* 1960, *20,* 3–8. (b)

GHISELLI, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, *47,* 81–86.

GHISELLI, E. E. *The validity of occupational aptitude tests.* New York: Wiley, 1966.

GHISELLI, E. E. Interaction of traits and motivational factors in the determination of the success of managers. *Journal of Applied Psychology,* 1968, *52,* 480–483.

GHISELLI, E. E. Comment on the use of moderator variables. *Journal of Applied Psychology,* 1972, *56,* 270.

GHISELLI, E. E., & HAIRE, M. The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychology,* 1960, *13,* 225–231.

GHISELLI, E. E., & SANDERS, E. P. Moderating heteroscedasticity. *Educational and Psychological Measurement,* 1967, *27,* 581–590.

GILBERSTADT, H., & DUKER, J. *A handbook for clinical and actuarial MMPI interpretation.* Philadelphia: Saunders, 1965.

GILBERT, H. B. On the IQ ban. *Teachers College Record,* 1966, *67,* 282–285.

GILBERT, J. A. Researches on the mental and physical development of school children. *Studies from the Yale Psychological Laboratory,* 1894, *2,* 40–100.

GILLILAND, A. R., & COLGIN, R. Norms, reliability, and forms of the MMPI. *Journal of Consulting Psychology,* 1951, *15,* 435–438.

GINSBURG, H., & OPPER, S. *Piaget's theory of intellectual development: An introduction.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist,* 1963, *18,* 519–522.

Glaser, R. Adapting the elementary school curriculum to individual performance. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1968. Pp. 3–36.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971. Ch. 17.

Glasser, A. J., & Zimmerman, I. L. *Clinical interpretation of the Wechsler Intelligence Scale for Children.* New York: Grune & Stratton, 1967.

Goffeney, B., Henderson, N. B., & Butler, B. V. Negro-white, male-female eight-month developmental scores compared with seven-year WISC and Bender Test scores. *Child Development,* 1971, *42,* 595–604.

Goldberg, L. R. Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs,* 1965, *79*(9, Whole No. 602).

Goldberg, L. R. Seer over sign: the first "good" example? *Journal of Experimental Research in Personality,* 1968, *3,* 168–171. (a)

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, *23,* 483–496. (b)

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, *73,* 422–432.

Goldberg, L. R. A historical survey of personality scales and inventories. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 2. Palo Alto, Calif.: Science and Behavior Books, 1971. Pp. 293–336.

Golden, M. Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology,* 1964, *28,* 440–446.

Goldfried, M. R., & D'Zurilla, T. J. A behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed.), *Current topics in clinical psychology.* Vol. 1. New York: Academic Press, 1969. Pp. 151–196.

Goldfried, M. R., & Kent, R. N. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin,* 1972, *77,* 409–420.

Goldfried, M. R., & Sprafkin, J. N. Behavioral personality assessment. *University Programs Modular Studies.* Morristown, N.J.: General Learning Corporation, 1974.

Goldfried, M. R., Stricker, G., & Weiner, I. R. *Rorschach handbook of clinical and research applications.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Golding, S. L., & Rorer, L. G. Illusory correlation and subjective judgment. *Journal of Abnormal Psychology,* 1972, *80,* 249–260.

Goldman, L. *Using tests in counseling.* (2nd ed.) New York: Appleton-Century-Crofts, 1971.

Goldschmid, M. L. Role of experience in the acquisition of conservation. *Proceedings, 76th Annual Convention, American Psychological Association,* 1968, 361–362.

Goldschmid, M. L., et al. A cross-cultural investigation of conservation. *Journal of Cross-Cultural Psychology,* 1973, *4,* 75–88.

GOLDSCHMID, M. L., & BENTLER, P. M. Dimensions and measurement of conservation. *Child Development,* 1968, *39,* 787–802. (a)

GOLDSCHMID, M. L., & BENTLER, P. M. *Manual: Concept Assessment Kit—Conservation.* San Diego, Calif.: Educational and Industrial Testing Service, 1968. (b)

GOLDSMITH, J. A. Systematic development and evaluation of a behavioral program for training psychiatric inpatients in interpersonal skills. Unpublished doctoral dissertation, University of Wisconsin, 1973.

GOODENOUGH, F. L. *Mental testing: Its history, principles, and applications.* New York: Rinehart, 1949.

GOODENOUGH, F. L., & HARRIS, D. B. Studies in the psychology of children's drawings: II. 1928–1949. *Psychological Bulletin,* 1950, *47,* 369–433.

GORDON, E. The Music Aptitude Profile: A new and unique musical aptitude test battery. *Council for Research in Music Education,* 1965, No. 6, 12–16.

GORDON, E. *A three-year longitudinal predictive validity study of the Musical Aptitude Profile.* Iowa City, Iowa: University of Iowa Press, 1967.

GORDON, E. W., & WILKERSON, D. A. *Compensatory education for the disadvantaged—programs and practices: Preschool through college.* New York: College Entrance Examination Board, 1966.

GORDON, L. V. *The measurement of interpersonal values.* Chicago: Science Research Associates, 1975.

GORDON, L. V., & ALF, E. F. Acclimatization and aptitude test performance. *Educational and Psychological Measurement,* 1960, *20,* 333–337.

GORDON, M. A. A study of the applicability of the same minimum qualifying scores for technical schools to white males, WAF, and Negro males. Lackland AFB, Texas: Human Resources Research Center. Technical Report 53–54, 1953.

GORHAM, D. R. Validity and reliability studies of a computer-based scoring system for inkblot responses. *Journal of Consulting Psychology,* 1967, *31,* 65–70.

GOTTFRIED, A. W., & BRODY, N. Interrelationships between and correlates of psychometric and Piagetian scales of sensorimotor intelligence. *Developmental Psychology,* 1975, *11,* 379–387.

GOUGH, H. G. A new dimension of status. I. Development of a personality scale. *American Sociological Review,* 1948, *13,* 401–409.

GOUGH, H. G. The Adjective Check List as a personality assessment research technique. *Psychological Reports,* 1960, *6,* 107–122.

GOUGH, H. G. A cluster analysis of Home Index status items. *Psychological Reports,* 1971, *28,* 923–929. (a)

GOUGH, H. G. Socioeconomic status as related to high school graduation and college attendance. *Psychology in the Schools,* 1971, *7,* 226–231. (b)

GOUGH, H. G., & HEILBRUN, A. B., JR. *The Adjective Check List manual.* Palo Alto, Calif.: Consulting Psychologists Press, 1972.

GOULET, L. R., & BALTES, P. B. (Eds.) *Life-span developmental psychology: Research and theory.* New York: Academic Press, 1970.

GRAHAM, F. K., et al. Development three years after perinatal anoxia and other potentially damaging newborn experiences. *Psychological Monographs,* 1962, *76*(3, Whole No. 522).

GRAHAM, F. K., ERNHART, C. B., et al. Brain injury in the preschool child. *Psychological Monographs,* 1963, *77*(10–11, Whole Nos. 573–574).

GRANT, D. L., & BRAY, D. W. Contributions of the interview to assessment of management potential. *Journal of Applied Psychology,* 1969, *53,* 24–34.

GRANT, D. L., & BRAY, D. W. Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology,* 1970, *54,* 7–14.

GREEN, B. F., JR. Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970. Pp. 184–197.

GREEN, D. R. (Ed.) *The aptitude-achievement distinction: Proceedings of the Second CTB/McGraw-Hill Conference on Issues in Educational Measurement.* New York: McGraw-Hill, 1974.

GREEN, D. R., FORD, M. P., & FLAMER, G. B. (Eds.) *Measurement and Piaget: Proceedings of the CTB/McGraw-Hill Conference on Ordinal Scales of Cognitive Development.* New York: McGraw-Hill, 1971.

GREEN, R. F. Does a selection situation induce testees to bias their answers on interest and temperament tests? *Educational and Psychological Measurement,* 1951, *11,* 503–515.

GREENWOOD, J. M., & MCNAMARA, W. J. Interrater reliability in situational tests. *Journal of Applied Psychology,* 1967, *51,* 101–106.

GRONLUND, N. E. *Stating behavioral objectives for classroom instruction.* New York: Macmillan, 1970.

GRONLUND, N. E. *Preparing criterion-referenced tests for classroom instruction.* New York: Macmillan, 1973.

GRONLUND, N. E. *Determining accountability for classroom instruction.* New York: Macmillan, 1974.

GRONLUND, N. E. *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan, 1976.

GROOMS, R. R., & ENDLER, N. S. The effect of anxiety on academic achievement. *Journal of Educational Psychology,* 1960, *51,* 299–304.

GROSS, A. L., & SU, W. H. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology,* 1975, *60,* 345–351.

GROSSMAN, H. (Ed.) *Manual on terminology and classification in mental retardation* (1973 revision). Washington, D.C. American Association on Mental Deficiency, 1973.

GUERTIN, W. H., et al. Research with the Wechsler-Bellevue Intelligence Scale: 1950–1955. *Psychological Bulletin,* 1956, *53,* 235–257.

GUERTIN, W. H., et al. Research with the Wechsler Intelligence Scales for Adults: 1955–1960. *Psychological Bulletin,* 1962, *59,* 1–26.

GUERTIN, W. H., et al. Research with the Wechsler Intelligence Scales for Adults: 1960–1965. *Psychological Bulletin,* 1966, *66,* 385–409.

GUERTIN, W. H., et al. Research with the Wechsler Intelligence Scales for Adults: 1965–1970. *Psychological Record,* 1971, *21,* 289–339.

GUICCIARDI, G., & FERRARI, G. C. I testi mentali per l'esame degli alienati. *Rivista sperimentale di Freniatria,* 1896, *22,* 297–314.

Guide to the use of the Graduate Record Examinations, 1973–74. Princeton, N.J.: Educational Testing Service, 1973.

GUILFORD, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

GUILFORD, J. P. Creative abilities in the arts. *Psychological Review,* 1957, *64,* 110–118.

GUILFORD, J. P. *Personality.* New York: Mc-Graw-Hill, 1959.

GUILFORD, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

GUILFORD, J. P., et al. A factor analysis study of human interests. *Psychological Monographs,* 1954, *68*(4, Whole No. 375).

GUILFORD, J. P., & FRUCHTER, B. *Fundamental statistics in psychology and education.* (5th ed.) New York: McGraw-Hill, 1973.

GUILFORD, J. P., & HOEPFNER, R. *The analysis of intelligence.* New York: McGraw-Hill, 1971.

GUILFORD, J. P., & LACEY, J. I. (Eds.) *Printed classification tests.* (AAF Aviation Psychology Program, Research Reports. Rep. No. 5.) Washington, D.C.: Government Printing Office, 1947.

GUILFORD, J. P., & ZIMMERMAN, W. S. Fourteen dimensions of temperament. *Psychological Monographs,* 1956, *70*(10, Whole No. 417).

GUION, R. M. Synthetic validity in a small company: A demonstration. *Personnel Psychology,* 1965, *18,* 49–63.

GULLIKSEN, H. The reliability of speeded tests. *Psychometrika,* 1950, *15,* 259–269. (a)

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950. (b)

GULLIKSEN, H. Looking back and ahead in psychometrics. *American Psychologist,* 1974, *29,* 251–261.

GULLIKSEN, H., & WILKS, S. S. Regression tests for several samples. *Psychometrika,* 1950, *15,* 91–114.

GURIN, P., GURIN, G., LAO, R., & BEATTIE, M. Internal-external control in the motivational dynamics of Negro youth. *Journal of Social Issues,* 1969, *25,* 29–53.

GUTTMAN, I., & RAJU, N. S. A minimum loss function as determiner of optimal cutting scores. *Personnel Psychology,* 1965, *18,* 179–185.

GUTTMAN, L. A basis for scaling qualitative ideas. *American Sociological Review,* 1944, *9,* 139–150.

GUTTMAN, L. Reliability formulas for noncompleted or speeded tests. *Psychometrika,* 1955, *20,* 113–124.

HAAN, N. Proposed model of ego functioning: Coping and defense mechanisms in relationship to IQ change. *Psychological Monographs,* 1963, *77*(8, Whole No. 571).

HALL, J. A correlation of a modified form of Raven's Progressive Matrices (1938) with Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology,* 1957, *21,* 23–26.

HALVERSON, H. M. The acquisition of skill in infancy. *Journal of Genetic Psychology,* 1933, *43,* 3–48.

HAMBLETON, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research,* 1974, *44,* 371–400.

HAMBLETON, R. K., & NOVICK, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement,* 1973, *10,* 159–170.

HAMILTON, R. G., & ROBERTSON, M. H. Examiner influence on the Holtzman Inkblot Technique. *Journal of Projective Techniques and Personality Assessment,* 1966, *30,* 553–558.

HAMMOND, K. R., & ALLEN, J. M. *Writing clinical reports.* Englewood Cliffs, N.J.: Prentice-Hall, 1953.

HAMMOND, K. R., HURSCH, C. J., & TODD, F. J. Analyzing the components of clinical inference. *Psychological Review,* 1964, *71,* 438–456.

HANSEN, J.–I. C., & JOHANSSON, C. B. Strong Vocational Interest Blank and dogmatism. *Journal of Counseling Psychology,* 1974, *21,* 196–201.

HARARI, O., & ZEDECK, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology,* 1973, *58,* 261–265.

HARCOURT BRACE JOVANOVICH. Selected items revised or eliminated from the Metropolitan as a result of possible bias against minority groups. *Metropolitan Achievement Tests Special Report* No. 23, 1972.

HARLOW, H. F. The formation of learning sets. *Psychological Review,* 1949, *56,* 51–65.

HARLOW, H. F. Learning set and error factor theory. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 2. New York: McGraw-Hill, 1960. Pp. 492–537.

HARMAN, H. H. *Modern factor analysis.* (2nd ed.) Chicago: University of Chicago Press, 1967.

HÄRNQVIST, K. Relative changes in intelligence from 13 to 18. I. Background and methodology. II. Results. *Scandinavian Journal of Psychology,* 1968, *9,* 50–82.

HARRIS, C. W., & MCARTHUR, D. L. Another view of the relation of environment to mental abilities. *Journal of Educational Psychology,* 1974, *66,* 457–459.

HARRIS, D. B. *Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough Draw-a-Man Test.* New York: Harcourt, Brace and World, 1963.

HARRIS, J. A. The computer: Guidance tool of the future. In W. E. Coffman (Ed.), *Frontiers of educational measurement and information systems—1973.* Boston: Houghton Mifflin, 1973. Ch. 7.

HARRISON, R. Thematic apperception methods. In B. B. Wolman (Ed.), *Handbook of clinical psychology.* New York: McGraw-Hill, 1965. Pp. 562–620.

HARRISON, R. H., & KASS, E. H. Differences between Negro and white pregnant women on the MMPI. *Journal of Consulting Psychology,* 1967, *31,* 454–463.

HARTSHORNE, H., & MAY, M. A. *Studies in deceit.* New York: Macmillan, 1928.

HARTSHORNE, H., MAY, M. A., & MALLER, J. B. *Studies in service and self-control.* New York: Macmillan, 1929.

HARTSHORNE, H., MAY, M. A., & SHUTTLEWORTH, F. K. *Studies in the organization of character.* New York: Macmillan, 1930.

HASE, H. D., & GOLDBERG, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin,* 1967, *67,* 231–248.

Hathaway, S. R., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory: Manual for administration and scoring.* New York: Psychological Corporation, 1967.

Hathaway, S. R., & Meehl, P. E. *An atlas for the clinical use of the MMPI.* Minneapolis: University of Minnesota Press, 1951.

Hathaway, S. R., & Monachesi, E. D. *Analyzing and predicting juvenile delinquency with the MMPI.* Minneapolis: University of Minnesota Press, 1953.

Hathaway, S. R., & Monachesi, E. D. *An atlas of juvenile MMPI profiles.* Minneapolis: University of Minnesota Press, 1961.

Hathaway, S. R., & Monachesi, E. D. *Adolescent personality and behavior.* Minneapolis: University of Minnesota Press, 1963.

Havighurst, R. J. *Human development and education.* New York: Longmans, Green, 1953.

Hay, E. N. Comparative validities in clerical testing. *Journal of Applied Psychology,* 1954, *38,* 299–301.

Hayes, K. J. Genes, drives, and intellect. *Psychological Reports,* 1962, *10,* 299–342.

Hayes, S. P. Alternative scales for the mental measurement of the visually handicapped. *Outlook for the Blind,* 1942, *36,* 225–230.

Hayes, S. P. A second test scale for the mental measurement of the visually handicapped. *Outlook for the Blind,* 1943, *37,* 37–41.

Hebb, D. O. Heredity and environment in mammalian behavior. *British Journal of Animal Behavior,* 1953, *1,* 43–47.

Hebb, D. O. A return to Jensen and his social science critics. *American Psychologist,* 1970, *25,* 568.

Heilbrun, A. B. Social-learning theory, social desirability, and the MMPI. *Psychological Bulletin,* 1964, *61,* 377–387.

Heim, A. W., & Wallace, J. G. The effects of repeatedly retesting the same group on the same intelligence test. *Quarterly Journal of Experimental Psychology,* 1949, *1,* 151–159; 1950, *2,* 19–32.

Helmreich, R., Bakeman, R., & Radloff, R. The Life History Questionnaire as a predictor of performance in Navy divers training. *Journal of Applied Psychology,* 1973, *57,* 148–153.

Helmstadter, G. C., & Ortmeyer, D. H. Some techniques for determining the relative magnitude of speed and power components of a test. *Educational and Psychological Measurement,* 1953, *13,* 280–287.

Henry, W. E. *The analysis of fantasy: The thematic apperception technique in the study of personality.* New York: Wiley, 1956.

Henry, W. E., & Farley, J. The validity of the Thematic Apperception Test in the study of adolescent personality. *Psychological Monographs,* 1959, *73*(17, Whole No. 487).

Henrysson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971. Ch. 5.

Herron, E. W. Psychometric characteristics of a thirty-item version of the group method of the Holtzman Inkblot Technique. *Journal of Clinical Psychology,* 1963, *19,* 450–453.

Herron, E. W. Changes in inkblot perception with presentation of the Holtzman Inkblot Technique as an "intelligence test." *Journal of Projective Techniques and Personality Assessment,* 1964, *28,* 442–447.

Herzberg, F., & Bouton, A. A further study of the stability of the Kuder Preference Record. *Educational and Psychological Measurement,* 1954, *14,* 326–331.

Hewer, V. H. Are tests fair to college students from homes with low socioeconomic status? *Personnel and Guidance Journal,* 1965, *43,* 764–769.

Hicks, L. E. Some properties of ipsative, normative, and forced normative measures. *Psychological Bulletin,* 1970, *74,* 167–184.

Higgins, C., and Sivers, C. H. A comparison of Stanford-Binet and Coloured Raven Progressive Matrices IQs for children with low socioeconomic status. *Journal of Consulting Psychology,* 1958, *22,* 465–468.

Hill, E. F. *The Holtzman Inkblot Technique: A handbook for clinical application.* San Francisco: Jossey-Bass, 1972.

Hill, K. T., & Sarason, S. B. The relation of test anxiety and defensiveness to test and school performance over the elementary school years. *Monographs of the Society for Research in Child Development,* 1966, *31*(2, Serial No. 104).

Himelstein, P. Research with the Stanford-Binet, Form L-M: The first five years. *Psychological Bulletin,* 1966, *65,* 156–164.

Hinckley, E. D. A follow-up study on the influence of individual opinion on the construction of an attitude scale. *Journal of Abnormal and Social Psychology,* 1963, *67,* 290–292.

Hobert, R., & Dunnette, M. D. Development of moderator variables to enhance the prediction of managerial effectiveness. *Journal of Applied Psychology,* 1967, *51,* 50–64.

Hoffman, B. *The tyranny of testing.* New York: Crowell-Collier, 1962.

Holden, R. H. Improved methods in testing cerebral palsied children. *American Journal of Mental Deficiency,* 1951, *56,* 349–353.

Holland, J. L. *The psychology of vocational choice.* Waltham, Mass.: Blaisdell, 1966.

Holland, J. L. *Making vocational choices: A theory of careers.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Hollander, E. P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology,* 1965, *49,* 434–438.

Hollenbeck, G. P., & Kaufman, A. S. Factor analysis of the Wechsler Preschool and Primary Scale of Intelligence (WPSSI). *Journal of Clinical Psychology,* 1973, *29,* 41–45.

Hollingshead, A. B. Two-factor index of social position. Unpublished manuscript, 1957. (Available from A. B. Hollingshead, Department of Sociology, Yale University, New Haven, Connecticut 06510.)

Holmen, M. G., et al. An assessment program for OCS applicants. *HumRRO Technical Report No. 26,* 1956.

Holt, R. R. *Assessing personality.* New York: Harcourt Brace Jovanovich, 1971.

Holtzman, W. H. Holtzman Inkblot Technique. In A. I. Rabin (Ed.), *Pro-*

jective techniques in personality assessment. New York: Springer, 1968. Pp. 136–170.

HOLTZMAN, W. H. (Ed.) *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

HOLTZMAN, W. H. The changing world of mental measurement and its social significance. *American Psychologist,* 1971, *26,* 546–553.

HOLTZMAN, W. H. (Chmn.) Moral development. *Proceedings of the 1974 Invitational Conference, Educational Testing Service,* 1975. (a)

HOLTZMAN, W. H. New developments in Holtzman Inkblot Technique. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975. Ch. 6. (b)

HOLTZMAN, W. H., MOSELEY, E. C., REINEHR, R. C., & ABBOTT, E. Comparison of the group method and the standard individual version of the Holtzman Inkblot Technique. *Journal of Clinical Psychology,* 1963, *19,* 441–449.

HOLTZMAN, W. H., THORPE, J. S., SWARTZ, J. D., & HERRON, E. W. *Inkblot perception and personality—Holtzman Inkblot Technique.* Austin, Texas: University of Texas Press, 1961.

HOLZBERG, J. D. Projective techniques. In I. A. Berg & L. A. Pennington (Eds.), *An introduction to clinical psychology.* New York: Ronald, 1966. Pp. 106–153.

HONZIK, M. P. Environmental correlates of mental growth: Prediction from the family setting at 21 months. *Child Development,* 1967, *38,* 337–364.

HONZIK, M. P., MACFARLANE, J. W., & ALLEN, L. The stability of mental test performance between two and eighteen years. *Journal of Experimental Education,* 1948, *17,* 309–324.

HOOPER, F. H. Cognitive assessment across the life-span: Methodological implications of the organismic approach. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues.* New York: Academic Press, 1973. Pp. 299–316.

HORN, C. C., & SMITH, L. F. The Horn Art Aptitude Inventory. *Journal of Applied Psychology,* 1945, *29,* 350–355.

HORST, P. A technique for the development of a differential prediction battery. *Psychological Monographs,* 1954, *68*(9, Whole No. 380).

HOVLAND, C. I., & SHERIF, M. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *Journal of Abnormal and Social Psychology,* 1952, *47,* 822–832.

HOWARD, J. L., & PLANT, W. T. Psychometric evaluation of an Operation Headstart program. *Journal of Genetic Psychology,* 1967, *111,* 281–288.

HUBER, J. T. *Report writing in psychology and psychiatry.* New York: Harper, 1961.

HUDDLESTON, E. M. Test development on the basis of content validity. *Educational and Psychological Measurement,* 1956, *16,* 283–293.

HUGHES, R. B., & LESSLER, K. A comparison of WISC and Peabody scores of Negro and white rural school children. *American Journal of Mental Deficiency,* 1965, *69,* 877–880.

HULL, C. L. *Aptitude testing.* Yonkers, N.Y.: World Book Co., 1928.

HUMPHREYS, L. G. The organization of human abilities. *American Psychologist,* 1962, *17,* 475–483.

HUMPHREYS, L. G. A skeptical look at the factor pure test. In C. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology.* Seattle, Washington: University of Washington, 1970. Pp. 23–32.

HUNT, J. McV., & KIRK, G. E. Criterion-referenced tests of school readiness: A paradigm with illustrations. *Genetic Psychology Monographs,* 1974, *90,* 143–182.

HUNT, J. McV., PARASKEVOPOULOS, J., SCHICKEDANZ, D., & UZGIRIS, I. C. Variations in the mean ages of achieving object permanence under diverse conditions of rearing. In B. Z. Friedlander, G. M. Sterritt, & G. E. Kirk (Eds.), *The exceptional infant: Assessment and intervention.* Vol. 3. New York: Brunner/Mazel, 1975. Pp. 247–263.

HUSÉN, T. The influence of schooling upon IQ. *Theoria,* 1951, *17,* 61–88.

ILG, F. L., & AMES, L. B. *School readiness: Behavior tests used at the Gesell Institute.* New York: Harper & Row, 1965.

ILG, F. L., AMES, L. B., & APELL, R. J. School readiness as evaluated by Gesell developmental, visual, and projective tests. *Genetic Psychology Monographs,* 1965, *71,* 61–91.

IRETON, H., THWING, E., & GRAVEM, H. Infant mental development and neurological status, family socioeconomic status, and intelligence at age four. *Child Development,* 1970, *41,* 937–945.

IRVINE, S. H. Factor analysis of African abilities and attainments: Constructs across cultures. *Psychological Bulletin,* 1969, *71,* 20–32. (a)

IRVINE, S. H. Figural tests of reasoning in Africa: Studies in the use of Raven's Matrices across cultures. *International Journal of Psychology,* 1969, *4,* 217–228. (b)

JACKSON, D. N. Balanced scales, item overlap, and the stables of Augeas. *Educational and Psychological Measurement,* 1967, *27,* 502–507. (a)

JACKSON, D. N. Block's The challenge of response sets. *Educational and Psychological Measurement,* 1967, *27,* 207–219. (b)

JACKSON, D. N. *Personality Research Form manual.* Goshen, N.Y.: Research Psychologists Press, 1967. (c)

JACKSON, D. N. A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology.* Vol. 2. New York: Academic Press, 1970. Pp. 61–96.

JACKSON, D. N. The dynamics of structured personality tests. *Psychological Review,* 1971, *78,* 229–248.

JACKSON, D. N. Structured personality assessment. In B. B. Wolman (Ed.), *Handbook of general psychology.* Englewood Cliffs, N.J.: Prentice-Hall, 1973. Pp. 775–792.

JACKSON, D. N., & MESSICK, S. Content and style in personality assessment. *Psychological Bulletin,* 1958, *55,* 243–252.

JACKSON, D. N., & MESSICK, S. Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition.* New York: Wiley, 1962. Pp. 129–155.

JACOBS, A., & BARRON, R. Falsification of the Guilford-Zimmerman Temperament Survey: II. Making a poor impression. *Psychological Reports,* 1968, *23,* 1271–1277.

JAEGER, R. M. The national test-equating study in reading (The Anchor Test Study). *NCME Measurement in Education,* 1973, *4*(4), 1–8.

JAMES, L. R. Criterion models and construct validity for criteria. *Psychological Bulletin,* 1973, *80,* 75–83.

JAMES, L. R., ELLISON, R. L., FOX, D. G., & TAYLOR, C. W. Prediction of artistic performance from biographical data. *Journal of Applied Psychology,* 1974, *59,* 84–86.

JANKE, L. L., & HAVIGHURST, R. J. Relations between ability and social status in a midwestern community: II. Sixteen-year-old boys and girls. *Journal of Educational Psychology,* 1945, *36,* 499–509.

JARRETT, R. F. Percent increase in output of selected personnel as an index of test efficiency. *Journal of Applied Psychology,* 1948, *32,* 135–146.

JENKINS, J. J., & RUSSELL, W. A. Systematic changes in word association norms: 1910–1952. *Journal of Abnormal and Social Psychology,* 1960, *60,* 293–304.

JENSEN, A. R. Social class and verbal learning. In M. Deutsch, I. Katz, & A. R. Jensen (Eds.), *Social class, race, and psychological development.* New York: Holt, Rinehart and Winston, 1968. Ch. 4.

JENSEN, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review,* 1969, *39,* 1–123.

JOHNSON, A. P. Notes on a suggested index of item validity: The U-L index. *Journal of Educational Psychology,* 1951, *42,* 499–504.

JOHNSON, D. J., & MYKLEBUST, H. R. *Learning disabilities: Educational principles and practices.* New York: Grune & Stratton, 1967.

JOHNSON, O. G., & BOMMARITO, J. W. *Tests and measurements in child development: A handbook.* San Francisco: Jossey-Bass, 1971.

JONES, H. E., CONRAD, H. S., & BLANCHARD, M. B. Environmental handicap in mental test performance. *University of California Publications in Psychology,* 1932, *5,* No. 3, 63–99.

JONES, H. G. The evaluation of the significance of differences between scaled scores on the WAIS: Perpetuation of a fallacy. *Journal of Consulting Psychology,* 1956, *20,* 319–320.

JONES, L. V. A factor analysis of the Stanford-Binet at four age levels. *Psychometrika,* 1949, *14,* 299–331.

JONES, L. V. Primary abilities in the Stanford-Binet, age 13. *Journal of Genetic Psychology,* 1954, *84,* 125–147.

JONES, M. C. Albert, Peter, and John B. Watson. *American Psychologist,* 1974, *29,* 581–583.

JONES, R. R., REID, J. B., & PATTERSON, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975. Ch. 2.

JUNG, C. G. The association method. *American Journal of Psychology,* 1910, *21,* 219–269.

JURGENSEN, C. E. Report on the "classification inventory," a personality test for industrial use. *Journal of Applied Psychology,* 1944, *28,* 445–460.

JURGENSEN, C. E. Table for determining phi coefficients. *Psychometrika,* 1947, *12,* 17–29.

KAGAN, J., & FREEMAN, M. Relation of childhood intelligence, maternal behaviors, and social class to behavior during adolescence. *Child Development,* 1963, *34,* 899–911.

KAGAN, J., SONTAG, L. W., BAKER, C. T., & NELSON, V. L. Personality and

IQ change. *Journal of Abnormal and Social Psychology,* 1958, *56,* 261–266.

KAHN, R. L. & CANNELL, C. F. *The dynamics of interviewing: Theory, technique, and cases.* New York: Wiley, 1957.

KAHNEMAN, D., & GHISELLI, E. E. Validity and nonlinear heteroscedastic models. *Personnel Psychology,* 1962, *15,* 1–11.

KAHNEMAN, D., & TVERSKY, A. On the psychology of prediction. *Psychological Review,* 1973, *80,* 237–251.

KAISER, H. F. A modified stanine scale. *Journal of Experimental Education,* 1958, *26,* 261.

KANFER, F. H. Assessment for behavior modification. *Journal of Personality Assessment,* 1972, *36,* 418–423.

KANFER, F. H., & SASLOW, G. Behavioral diagnosis. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status.* New York: McGraw-Hill, 1969. Pp. 417–444.

KAPLAN, L. J. On the testing of nontestable children. *Bulletin of the Menninger Clinic,* 1975, *39,* 420–425.

KATZ, E. The "Pointing Modification" of the Revised Stanford-Binet Intelligence Scales, Forms L and M, Years II through VI: A report of research in progress. *American Journal of Mental Deficiency,* 1958, *62,* 698–707.

KATZ, M. R. Interpreting Kuder Preference Record scores: Ipsative or normative. *Vocational Guidance Quarterly,* 1962, *10,* 96–100.

KATZ, M. R. Career decision-making: A computer-based system of interactive guidance and information (SIGI). *Proceedings of the 1973 Invitational Conference on Testing Problems, Educational Testing Service,* 1974, 43–69.

KAUFMAN, A. S. A short form of the Wechsler Preschool and Primary Scale of Intelligence. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 361–369.

KAUFMAN, A. S. Comparison of the WPPSI, Stanford-Binet, and McCarthy scales as predictors of first-grade achievement. *Perceptual and Motor Skills,* 1973, *36,* 67–73. (a)

KAUFMAN, A. S. The relationship of WPPSI IQs to SES and other background variables. *Journal of Clinical Psychology,* 1973, *29,* 354–357. (b)

KAUFMAN, A. S. Factor analysis of the WISC-R at eleven age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology,* 1975, *43,* 135–147. (a)

KAUFMAN, A. S. Factor structure of the McCarthy scales at five age levels between 2½ and 8½. *Educational and Psychological Measurement,* 1975, *35,* 641–656. (b)

KAUFMAN, A. S., & HOLLENBECK, G. P. Factor analysis of the standardization edition of the McCarthy scales. *Journal of Clinical Psychology,* 1973, *29,* 358–362.

KAUFMAN, A. S., & HOLLENBECK, G. P. Comparative structure of the WPPSI for blacks and whites. *Journal of Clinical Psychology,* 1974, *30,* 316–319.

KAUFMAN, A. S., & KAUFMAN, N. L. Tests built from Piaget's and Gesell's tasks as predictors of first-grade achievement. *Child Development,* 1972, *43,* 521–535.

KAUFMAN, A. S., & KAUFMAN, N. L. Black-white differences at ages 2½–8½ on the McCarthy Scales of Children's Abilities. *Journal of School Psychology,* 1973, *11,* 196–206. (a)

Kaufman, A. S., & Kaufman, N. L. Sex differences on the McCarthy Scales of Children's Abilities. *Journal of Clinical Psychology,* 1973, *29,* 362–365. (b)

Kaufman, A. S., & Kaufman, N. L. Social class differences on the McCarthy scales for black and white children. *Perceptual and Motor Skills,* 1975, *41,* 205–206.

Kaufman, N. L. & Kaufman, A. S. Comparison of normal and minimally brain dysfunctioned children on the McCarthy Scales of Children's Abilities. *Journal of Clinical Psychology,* 1974, *30,* 69–72.

Kavruck, S. Thirty-three years of test research: A short history of test development in the U.S. Civil Service Commission. *American Psychologist,* 1956, *11,* 329–333.

Kelley, T. L. *Interpretation of educational measurements.* Yonkers, N.Y.: World Book Co., 1927.

Kelley, T. L. *Crossroads in the mind of man: A study of differentiable mental abilities.* Stanford, Calif.: Stanford University Press, 1928.

Kelley, T. L. *Essential traits of mental life.* Cambridge, Mass.: Harvard University Press, 1935.

Kelley, T. L. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology,* 1939, *30,* 17–24.

Kelley, T. L. Cumulative significance of a number of independent experiments: Reply to A. E. Traxler and R. N. Hilkert. *School and Society,* 1943, *57,* 482–484.

Kelly, E. L. The place of situation tests in evaluating clinical psychologists. *Personnel Psychology,* 1954, *7,* 484–492.

Kelly, E. L., & Fiske, D. W. *The prediction of performance in clinical psychology.* Ann Arbor, Mich.: University of Michigan Press, 1951.

Kelly, G. A. *A theory of personality.* New York: Norton, 1963.

Kelly, G. A. A summary statement of a cognitively-oriented comprehensive theory of behavior. In J. C. Mancuso (Ed.), *Readings for a cognitive theory of personality.* New York: Holt, Rinehart and Winston, 1970. Pp. 27–58.

Kelman, H. C. Attitudes are alive and well and gainfully employed in the sphere of action. *American Psychologist,* 1974, *29,* 310–324.

Kendrick, S. A., & Thomas, C. L. Transition from school to college. *Review of Educational Research,* 1970, *40,* 151–179.

Kennedy, W. A., et al. The ceiling of the new Stanford-Binet. *Journal of Clinical Psychology,* 1960, *17,* 284–286.

Kennedy, W. A., Van de Reit, V., & White, J. C. A normative sample of intelligence and achievement of Negro elementary school children in the southeastern United States. *Monographs of the Society for Research in Child Development,* 1963, *28* (6, Serial No. 90).

Kent, G. H., & Rosanoff, A. J. A study of association in insanity. *American Journal of Insanity,* 1910, *67,* 37–96; 317–390.

King, W. L., & Seegmiller, B. Performance of 14- to 22-month-old black, firstborn male infants on two tests of cognitive development: The Bayley

Scales and the Infant Psychological Development Scale. *Developmental Psychology,* 1973, *8,* 317–326.

KINSLINGER, H. J. Application of projective techniques in personnel psychology since 1940. *Psychological Bulletin,* 1966, *66,* 134–149.

KINTER, M. The measurement of artistic abilities. New York: Psychological Corporation, 1933.

KIRCHNER, W. K. A note on the effect of privacy in taking typing tests. *Journal of Applied Psychology,* 1966, *50,* 373–374.

KIRK, S. A. *Educating exceptional children.* Boston: Houghton Mifflin, 1972.

KIRK, S. A., & KIRK, W. D. *Psycholinguistic learning disabilities: Diagnosis and remediation.* Urbana, Ill.: University of Illinois Press, 1971.

KIRKPATRICK, J. J., EWEN, R. B., BARRETT, R. S., & KATZELL, R. A. *Testing and fair employment.* New York: New York University Press, 1968.

KLAUS, R. A., & GRAY, S. W. The early training project for disadvantaged children: A report after five years. *Monographs of the Society for Research in Child Development,* 1968, *33*(4, Serial No. 120).

KLEINMUNTZ, B. *Personality measurement: An introduction.* Homewood, Ill.: Dorsey, 1967.

KLEINMUNTZ, B. The computer as clinician. *American Psychologist,* 1975, *30,* 379–387.

KLINEBERG, O. An experimental study of speed and other factors in "racial" differences. *Archives of Psychology,* 1928, No. 93.

KLINGER, E. Fantasy need achievement as a motivational construct. *Psychological Bulletin,* 1966, *66,* 291–308.

KNAPP, R. R. The effects of time limits on the intelligence test performance of Mexican and American subjects. *Journal of Educational Psychology,* 1960, *51,* 14–20.

KNOBLOCH, H., & PASAMANICK, B. An evaluation of the consistency and predictive value of the 40-week Gesell Developmental Schedule. *Psychiatric Research Reports,* 1960, *13,* 10–41.

KNOBLOCH, H., & PASAMANICK, B. Predicting intellectual potential in infancy. *American Journal of Diseases of Children,* 1963, *106,* 43–51.

KNOBLOCH, H., & PASAMANICK, B. Prospective studies on the epidemiology of reproductive casualty: Methods, findings, and some implications. *Merrill-Palmer Quarterly,* 1966, *12,* 27–43.

KNOBLOCH, H., & PASAMANICK, B. Prediction from the assessment of neuromotor and intellectual status in infancy. In *Psychopathology of mental development.* New York: Grune & Stratton, 1967.

KNOELL, D. M., & HARRIS, C. W. A factor analysis of spelling ability. *Journal of Educational Research,* 1952, *46,* 95–111.

KOCH, H. L. *Twins and twin relations.* Chicago: University of Chicago Press, 1966.

KOHLBERG, L. Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization: Theory and research.* Chicago: Rand McNally, 1969.

KOHLBERG, L. The development of moral stages: Uses and abuses. *Proceedings* *of the 1973 Invitational Conference on Testing Problems, Educational Testing Service,* 1974, 1–8.

KOPPITZ, E. M. *The Bender Gestalt Test for young children.* New York: Grune & Stratton, 1964.

KOPPITZ, E. M. *The Bender Gestalt Test for young children: Research and application, 1963–1973.* New York: Grune & Stratton, 1975.

KRAEPELIN, E. *Über die Beeinflüssung einfacher psychischer Vorgänge durch einige Arzneimittel.* Jena: Fischer, 1892.

KRAEPELIN, E. Der psychologische Versuch in der Psychiatrie. *Psychologische Arbeiten,* 1895, *1,* 1–91.

KRASNER, L. Behavior therapy. *Annual Review of Psychology,* 1971, *22,* 483–532.

KRATHWOHL, D. R., et al. *Taxonomy of educational objectives; the classification of educational goals; handbook 2: Affective domain.* New York: McKay, 1964.

KUDER, G. F. A rationale for evaluating interests. *Educational and Psychological Measurement,* 1963, *23,* 3–10.

KUDER, G. F. *Kuder General Interest Survey: Manual.* Chicago: Science Research Associates, 1964.

KUDER, G. F. *Kuder Occupational Interest Survey: General manual.* Chicago: Science Research Associates, 1966. (a)

KUDER, G. F. The occupational interest survey. *Personnel and Guidance Journal,* 1966, *45,* 72–77. (b)

KUDER, G. F. Some principles of interest measurement. *Educational and Psychological Measurement,* 1970, *30,* 205–226.

KUDER, G. F., & RICHARDSON, M. W. The theory of estimation of test reliability. *Psychometrika,* 1937, *2,* 151–160.

KUHLEN, R. G. Age and intelligence: The significance of cultural change in longitudinal vs. cross-sectional findings. *Vita Humana,* 1963, *6,* 113–124.

KUHLMANN, F. A revision of the Binet-Simon system for measuring the intelligence of children. *Journal of Psycho-Asthenics, Monograph Supplement,* 1912, *1,* 1–41.

KUNCE, J. T. Vocational interests and accident proneness. *Journal of Applied Psychology,* 1967, *51,* 223–225.

KURTINES, W., & GREIF, E. B. The development of moral thought: Review and evaluation of Kohlberg's approach. *Psychological Bulletin,* 1974, *81,* 453–470.

KURTZ, A. K. A research test of the Rorschach test. *Personnel Psychology,* 1948, *1,* 41–51.

LA FAVE, L. Essay vs. multiple-choice: Which test is preferable? *Psychology in the Schools,* 1966, *3,* 65–69.

LA FAVE, L. Attitude switching vs. attitude change: A neglected distinction. Paper presented at the meeting of the Canadian Psychological Association, Windsor, Ontario, June 1974.

LA FAVE, L., HADDAD, J., & MARSHALL, N. Humor judgments as a function of identification classes. *Sociology and Social Research,* 1974, *58,* 184–194.

LANYON, R. I. A free-choice version of the EPPS. *Journal of Clinical Psychology,* 1966, *22,* 202–205.

LANYON, R. I. *A handbook of MMPI group profiles.* Minneapolis: University of Minnesota Press, 1968.

LANYON, R. I., & GOODSTEIN, L. D. *Personality assessment.* New York: Wiley, 1971.

LAOSA, L. M., SWARTZ, J. D., & DIAZ-GUERRERO, R. Perceptual-cognitive and personality development of Mexican and Anglo-American children as measured by human figure drawings. *Developmental Psychology,* 1974, *10,* 131–139.

LARKIN, K. C., & WEISS, D. J. An empirical investigation of computer-administered pyramidal ability testing. Psychometrics Methods Program, Department of Psychology, University of Minnesota, Research Report 74–3, July 1974.

LARKIN, K. C., & WEISS, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Psychometric Methods Program, Department of Psychology, University of Minnesota, Research Report 75–1, February 1975.

LAUMANN, E. O., & HOUSE, J. S. Living room styles and social attributes: The patterning of material artifacts in a modern urban community. *Sociology and Social Research,* 1970, *54,* 321–342.

LAURENDEAU, M., & PINARD, A. *Causal thinking in the child: A genetic and experimental approach.* New York: International Universities Press, 1962.

LAURENDEAU, M., & PINARD, A. *The development of the concept of space in the child.* New York: International Universities Press, 1970.

LAWRENCE, S. W., JR. The effects of anxiety, achievement motivation, and task importance upon performance on an intelligence test. *Journal of Educational Psychology,* 1962, *53,* 150–156.

LAWSHE, C. H. Employee selection. *Personnel Psychology,* 1952, *5,* 31–34.

LAWSHE, C. H., & BALMA, M. J. *Principles of personnel testing.* (2nd ed.) New York: McGraw-Hill, 1966.

LEE, R., & BOOTH, J. M. A utility analysis of a weighted application blank designed to predict turnover for clerical employees. *Journal of Applied Psychology,* 1974, *59,* 516–518.

LEHMAN, P. R. *Tests and measurements in music.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

LENNON, R. T. Assumptions underlying the use of content validity. *Educational and Psychological Measurement,* 1956, *16,* 294–304.

LENNON, R. T. A comparison of results of three intelligence tests. In C. I. Chase & H. G. Ludlow (Eds.), *Readings in educational and psychological measurement.* Boston: Houghton Mifflin, 1966. Pp. 198–205. (a)

LENNON, R. T. Norms: 1963. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 243–250. (b)

LERNER, J. W. *Children with learning disabilities: Theories, diagnosis, and teaching strategies.* Boston: Houghton Mifflin, 1971.

LEVINSON, B. M. Traditional Jewish cultural values and performance on the Wechsler tests. *Journal of Educational Psychology,* 1959, *50,* 177–181.

LEVINSON, B. M. Sub-cultural values and IQ stability. *Journal of Genetic Psychology,* 1961, *98,* 69–82.

LEVINSON, B. M. The WAIS quotient of subcultural deviation. *Journal of Genetic Psychology,* 1963, *103,* 123–131.

LEVITT, E. E., & TRUUMAA, A. *The Rorschach technique with children and adolescents: Application and norms.* New York: Grune & Stratton, 1972.

LEVONIAN, E. A statistical analysis of the 16 Personality Factor Questionnaire. *Educational and Psychological Measurement,* 1961, *21,* 589–596.

LEVY, P. Short-form tests: A methodological review. *Psychological Bulletin,* 1968, *69,* 410–416.

LEWIS, M. Infant intelligence tests: Their use and misuse. *Human Development,* 1973, *16,* 108–118.

LEWIS, M., & MCGURK, H. Evaluation of infant intelligence: Infant intelligence scores—true or false? *Science,* 1972, *178*(4066), 1174–1177.

LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology,* 1932, No. 140.

LINDGREN, B. W., & MCELRATH, G. W. *Introduction to probability and statistics.* (3rd ed.) New York: Macmillan, 1969.

LINDSLEY, D. B. The psychology of lie detection. In G. J. Dudycha et al. *Psychology for law enforcement officers.* Springfield, Ill.: Charles C Thomas, 1955. Ch. 4.

LINDZEY, G. On the classification of projective techniques. *Psychological Bulletin,* 1959, *56,* 158–168.

LINDZEY, G. *Projective techniques and cross-cultural research.* New York: Appleton-Century-Crofts, 1961.

LINDZEY, G., & BORGATTA, E. F. Sociometric measurement. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954. Ch. 11.

LINDZEY, G., & HERMAN, P. S. Thematic Apperception Test: A note on reliability and situational validity. *Journal of Projective Techniques,* 1955, *19,* 36–42.

LINN, R. L. Fair test use in selection. *Review of Educational Research,* 1973, *43,* 139–161.

LINN, R. L. Test bias and the prediction of grades in law school. *Journal of Legal Education,* 1975, *27,* 293–323.

LINN, R. L., ROCK, D. A., & CLEARY, T. A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement,* 1969, *29,* 129–146.

LIPINSKI, D., & NELSON, R. O. The reactivity and unreliability of self-recording. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 118–123.

LITTELL, W. M. The Wechsler Intelligence Scale for Children: Review of a decade of research. *Psychological Bulletin,* 1960, *57,* 132–156.

LITTLE, K. B., & SHNEIDMAN, E. S. Congruencies among interpretations of psychological test and anamnestic data. *Psychological Monographs,* 1959, *73*(6, Whole No. 476).

LIVINGSTON, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement,* 1972, *9,* 13–26. (See also critiques and replies: 1972, *9,* 27–31; 1973, *10,* 49–50, 311.)

LOEHLIN, J., LINDZEY, G., & SPUHLER, J. N. *Race differences in intelligence.* San Francisco: Freeman, 1975.

LOEVINGER, J. The meaning and measurement of ego development. *American Psychologist,* 1966, *21,* 195–206. (a)

LOEVINGER, J. A theory of test response. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 545–556. (b)

LOEVINGER, J., & OSSORIO, A. G. Evaluation of therapy by self-report: A paradox. *American Psychologist,* 1958, *13,* 366.

LOEVINGER, J., WESSLER, R., & REDMORE, C. *Measuring ego development.* Vol. 1: *Construction and use of a sentence completion test.* Vol. 2: *Scoring manual for women and girls.* San Francisco: Jossey-Bass, 1970.

LONGSTAFF, H. P. Fakability of the Strong Interest Blank and the Kuder Preference Record. *Journal of Applied Psychology,* 1948, *32,* 360–369.

LONNER, W. J., & ADAMS, H. L. Interest patterns of psychologists in nine Western nations. *Journal of Applied Psychology,* 1972, *56,* 141–151.

LORD, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika,* 1952, *17,* 181–194.

LORD, F. M. Relation of test score to trait underlying the test. *Educational and Psychological Measurement,* 1953, *13,* 517–549.

LORD, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970. Pp. 139–183.

LORD, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement,* 1971, *8,* 147–151. (a)

LORD, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement,* 1971, *31,* 805–813. (b)

LORD, F. M. A theoretical study of two-stage testing. *Psychometrika,* 1971, *36,* 227–241. (c)

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

LORET, P. G., SEDER, A., BIANCHINI, J. C., & VALE, C. A. *Anchor Test Study: Equivalence and norms tables for selected reading achievement tests.* Washington, D.C.: Government Printing Office, 1974.

LORETAN, J. O. Alternatives to intelligence testing. *Proceedings of the 1965 Invitational Conference on Testing Problems, Educational Testing Service,* 1966, 19–30.

LORGE, I. Schooling makes a difference. *Teachers College Record,* 1945, *46,* 483–492.

LOTSOF, E. J. Intelligence, verbal fluency, and the Rorschach test. *Journal of Consulting Psychology,* 1953, *17,* 21–24.

LOVAAS, O. I., FREITAG, G., GOLD, V. J., & KASSORLA, I. C. Recording apparatus for observation of behaviors of children in free play settings. *Journal of Experimental Child Psychology,* 1965, *2,* 108–120.

LSAT handbook. Princeton, N.J.: Educational Testing Service, 1964.

LUBORSKY, L. B., & CATTELL, R. B. The validation of personality factors in humor. *Journal of Personality,* 1947, *15,* 283–291.

MACDONALD, M. A behavioral assessment methodology applied to the measurement of assertion. Unpublished doctoral dissertation, University of Illinois, 1974.

Machover, K. *Personality projection in the drawing of the human figure: A method of personality investigation.* Springfield, Ill.: Charles C Thomas, 1949.

MacKinney, A. C. The assessment of performance change: An inductive example. *Organizational Behavior and Human Performance,* 1967, *2,* 56–72.

MacPhillamy, D. J., & Lewinsohn, P. M. Depression as a function of levels of desired and obtained pleasure. *Journal of Abnormal Psychology,* 1974, *83,* 651-657.

MacPhillamy, D. J., & Lewinsohn, P. M. *Manual for the Pleasant Events Schedule.* Mimeo, 1975. (Available from P. M. Lewinsohn, Department of Psychology, University of Oregon, Eugene, Oregon 97403.)

Maier, M. H. Effects of educational level on prediction of training success with the ACB. U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 225, June 1972.

Maier, M. H., & Fuchs, E. F. Development and evaluation of a new ACB and aptitude area system. U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 239, September 1972.

Maier, M. H., & Fuchs, E. F. Effectiveness of selection and classification testing. U.S. Army Research Institute for the Behavioral and Social Sciences. Research Report 1179, September 1973.

Mallinson, G. G., & Crumrine, W. M. An investigation of the stability of interests of high school students. *Journal of Educational Research,* 1952, *45,* 369–383.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology,* 1952, *47,* 166–173.

Marjoribanks, K. Environment, social class, and mental abilities. *Journal of Educational Psychology,* 1972, *63,* 103–109.

Marjoribanks, K. Another view of the relation of environment to mental abilities. *Journal of Educational Psychology,* 1974, *66,* 460–463.

Marks, P. A., Seeman, W., & Haller, D. L. *The actuarial use of the MMPI with adolescents and adults.* Baltimore: Williams & Wilkins, 1974.

Masling, J. The effects of warm and cold interaction on the administration and scoring of an intelligence test. *Journal of Consulting Psychology,* 1959, *23,* 336–341.

Masling, J. The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin,* 1960, *57,* 65–85.

Masling, J. Differential indoctrination of examiners and Rorschach responses. *Journal of Consulting Psychology,* 1965, *29,* 198–201.

Masterson, S. The adjective checklist technique: A review and critique. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975. Ch. 7.

Matarazzo, J. D. *Wechsler's measurement and appraisal of adult intelligence.* Baltimore: Williams & Wilkins, 1972.

Matarazzo, J. D., & Wiens, A. N. *The interview: Research on its anatomy and structure.* Chicago: Aldine, 1972.

MAXFIELD, K. B., & BUCHHOLZ, S. *A social maturity scale for blind preschool children: A guide to its use.* New York: American Foundation for the Blind, 1957.

MAXWELL, E. Validities of abbreviated WAIS scales. *Journal of Consulting Psychology,* 1957, *21,* 121–126.

MCCALL, R. B., APPELBAUM, M. I., & HOGARTY, P. S. Developmental changes in mental performance. *Monographs of the Society for Research in Child Development,* 1973, *38*(3, Serial No. 150).

MCCALL, R. B., HOGARTY, P. S., & HURLBURT, N. Transitions in infant sensorimotor development and the prediction of childhood IQ. *American Psychologist,* 1972, *27,* 728–748.

MCCALL, W. A. *How to measure in education.* New York: Macmillan, 1922.

MCCARTHY, D. A study of the reliability of the Goodenough drawing test of intelligence. *Journal of Psychology,* 1944, *18,* 201–216.

MCCARTHY, J. J., & MCCARTHY, J. F. *Learning disabilities.* Boston: Allyn & Bacon, 1969.

MCCLELLAND, D. C. *The achieving society.* New York: Free Press, 1961.

MCCLELLAND, D. C. Longitudinal trends in the relation of thought to action. *Journal of Consulting Psychology,* 1966, *30,* 479–483.

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

MCCORMICK, E. J. Application of job analysis to indirect validity. *Personnel Psychology,* 1959, *12,* 395–420.

MCCORMICK, E. J., JEANNERET, P. R., & MECHAM, R. C. A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ). *Journal of Applied Psychology,* 1972, *56,* 347–368.

MCCORMICK, E. J., & TIFFIN, J. *Industrial psychology.* (6th ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1974.

MCDONALD, R. L., & GYNTHER, M. D. MMPI differences associated with sex, race, and class in two adolescent samples. *Journal of Consulting Psychology,* 1963, *27,* 112–116.

MCFALL, R. M., & LILLESAND, D. B. Behavior rehearsal with modeling and coaching in assertion training. *Journal of Abnormal Psychology,* 1971, *77,* 313–323.

MCFALL, R. M., & TWENTYMAN, C. T. Four experiments on the relative contributions of rehearsal, modeling, and coaching to assertion training. *Journal of Abnormal Psychology,* 1973, *81,* 199–218.

MCKECHNIE, G. E. *ERI manual: Environmental Response Inventory.* Palo Alto, Calif.: Consulting Psychologists Press, 1974.

MCLEISH, J. The validation of Seashore's measures of musical talent by factorial methods. *British Journal of Psychology, Statistical Section,* 1950, *3,* 129–140.

MCNEMAR, Q. *The revision of the Stanford-Binet Scale: An analysis of the standardization data.* Boston: Houghton Mifflin, 1942.

MCNEMAR, Q. On abbreviated Wechsler-Bellevue Scales. *Journal of Consulting Psychology,* 1950, *14,* 79–81.

MCREYNOLDS, P. (Ed.) *Advances in psychological assessment.* Palo Alto, Calif.: Science and Behavior Books. Vol. 1, 1968; Vol. 2, 1971.

McReynolds, P. (Ed.) *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975.

Meehl, P. E. An investigation of a general normality or control factor in personality testing. *Psychological Monographs,* 1945, 59(4, Whole No. 274).

Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press, 1954.

Meehl, P. E. Wanted—a good cookbook. *American Psychologist,* 1956, *11,* 263–272.

Meehl, P. E. When shall we use our heads instead of the formula? *Journal of Counseling Psychology,* 1957, *4,* 268–273.

Meehl, P. E. Seer over sign: The first good example. *Journal of Experimental Research in Personality,* 1965, *1,* 27–32.

Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin,* 1955, *52,* 194–216.

Meeker, M. N. *The structure of intellect: Its interpretation and uses.* Columbus, Ohio: Charles E. Merrill, 1969.

Megargee, E. I. The relation of response length to the Holtzman Inkblot Technique. *Journal of Consulting Psychology,* 1966, *30,* 415–419. (a)

Megargee, E. I. (Ed.) *Research in clinical assessment.* New York: Harper & Row, 1966. (b)

Megargee, E. I. *The California Psychological Inventory handbook.* San Francisco, Calif.: Jossey-Bass, 1972.

Megargee, E. I., & Parker, G. V. An exploration of the equivalence of Murrayan needs as assessed by the Adjective Check List, the TAT, and the Edwards Personal Preference Schedule. *Journal of Clinical Psychology,* 1968, *24,* 47–51.

Meier, N. C. *Art in human affairs.* New York: McGraw-Hill, 1942.

Melton, A. W. (Ed.) *Apparatus tests.* (AAF Aviation Psychology Program, Research Reports. Rep. No. 4) Washington, D.C.: Government Printing Office, 1947.

Merrens, M. R., & Richards, W. S. Acceptance of generalized versus "bona fide" personality interpretations. *Psychological Reports,* 1970, *27,* 691–694.

Messick, S. The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction: Issues and problems.* New York: Holt, Rinehart and Winston, 1970. Pp. 183–202.

Messick, S. The context of assessment and the assessment of context. In *Circus: Comprehensive assessment in nursery school and kindergarten.* (Proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973) Princeton, N.J.: Educational Testing Service, 1973. Pp. 33–38. (a)

Messick, S. Multivariate models of cognition and personality: The need for both process and structure in psychological theory and measurement. In

J. R. Royce (Ed.), *Multivariate analysis and psychological theory.* New York: Academic Press, 1973. Pp. 265–303. (b)

MESSICK, S., & KOGAN, N. Personality consistencies in judgment: Dimensions of role construct. *Multivariate Behavioral Research,* 1966, *1,* 165–175.

MIKLICH, D. R. Radio telemetry in clinical psychology and related areas. *American Psychologist,* 1975, *30,* 419–425.

MILGRAM, N. A. IQ constancy in disadvantaged Negro children. *Psychological Reports,* 1971, *29,* 319–326.

MILGRAM, N. A., & OZER, M. N. Peabody Picture Vocabulary Test scores of preschool children. *Psychological Reports,* 1967, *20,* 779–784.

MILLER, C., KNAPP, S. C., & DANIELS, C. W. MMPI study of Negro mental hygiene clinic patients. *Journal of Abnormal Psychology,* 1968, *73,* 168–173.

MILLER, R. J. Cross-cultural research in the perception of pictorial materials. *Psychological Bulletin,* 1973, *80,* 135–150.

MILLMAN, J. *Determining test length.* Los Angeles: Instructional Objectives Exchange, 1972.

MILLMAN, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research,* 1973, *43,* 205–216.

MILLMAN, J. Criterion-referenced measurement. In M. W. Apple & M. J. Subkoviak, *Educational evaluation: Analysis and responsibility.* Berkeley, Calif.: McCutchan, 1974. Ch. 6.

MILLMAN, J., BISHOP, C. H., & EBEL, R. An analysis of test-wiseness. *Educational and Psychological Measurement,* 1965, *25,* 707–726.

MISCHEL, W. *Personality and assessment.* New York: Wiley, 1968.

MISCHEL, W. Continuity and change in personality. *American Psychologist,* 1969, *24,* 1012–1018.

MISCHEL, W. Toward a cognitive social learning reconceptualization of personality. *Psychological Review,* 1973, *80,* 252–283.

MISCHEL, W., ZEISS, R., & ZEISS, A. Internal-external control and persistence: Validation and implications of the Stanford Preschool Internal-External Scale. *Journal of Personality and Social Psychology,* 1974, *29,* 265–278.

MITCHELL, B. C. Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for white and Negro pupils. *Educational and Psychological Measurement,* 1967, *27,* 1047–1054.

MOLLENKOPF, W. G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika,* 1950, *15,* 291–317. (a)

MOLLENKOPF, W. G. Predicted differences and differences between predictions. *Psychometrika,* 1950, *15,* 409–417. (b)

MOOS, R. H. Conceptualizations of human environments. *American Psychologist,* 1973, *28,* 652–663.

MOOS, R. H. *Evaluating treatment environments: A social ecological approach.* New York: Wiley, 1974. (a)

MOOS, R. H. *The Social Climate Scales: An overview.* Palo Alto, Calif.: Consulting Psychologists Press, 1974. (b)

Moos, R. H. Assessment and impact of social climate. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 3. San Francisco: Jossey-Bass, 1975. Ch. 1. (a)

Moos, R. H. *Evaluating correctional environments: With implications for community settings.* New York: Wiley, 1975. (b)

Moreno, J. L. *Who shall survive? Foundations of sociometry, group psychotherapy, and sociodrama.* (2nd ed.) Beacon, N.Y.: Beacon House, 1953. (*Sociometry Monographs,* No. 29.)

Moriarty, A. E. Children's ways of coping with the intelligence test. *Menninger Clinic Bulletin,* 1960, *24,* 115–127.

Moriarty, A. E. Coping patterns of preschool children in response to intelligence test demands. *Genetic Psychology Monographs,* 1961, *64,* 3–127.

Moriarty, A. E. *Constancy and IQ change: A clinical view of relationships between tested intelligence and personality.* Springfield, Ill.: Charles C Thomas, 1966.

Morrow, R. S. An analysis of the relations among tests of musical, artistic, and mechanical abilities. *Journal of Psychology,* 1938, *5,* 253–263.

Mosier, C. I. Problems and designs of cross-validation. *Educational and Psychological Measurement,* 1951, *11,* 5–11.

Mosier, C. I., & McQuitty, J. V. Methods of item validation and abacs for item-test correlation and critical ratio of upper-lower difference. *Psychometrika,* 1940, *5,* 57–65.

Mowrer, O. H. "Q-technique"—description, history, and critique. In O. H. Mowrer (Ed.), *Psychotherapy theory and research.* New York: Ronald, 1953. Pp. 316–375.

Mugaas, H. D., & Hester, R. The development of an equation for identifying the interests of carpenters. *Educational and Psychological Measurement,* 1952, *12,* 408–414.

Murphy, G., & Kovach, J. K. *Historical introduction to modern psychology.* (3rd ed.) New York: Harcourt Brace Jovanovich, 1972.

Murray, H. A. et al. *Explorations in personality.* New York: Oxford University Press, 1938.

Murray, H. A., & MacKinnon, D. W. Assessment of OSS personnel. *Journal of Consulting Psychology,* 1946, *10,* 76–80.

Murstein, B. I. *Theory and research in projective techniques (emphasizing the TAT).* New York: Wiley, 1963.

Mussen, P. H., & Naylor, H. K. The relationships between overt and fantasy aggression. *Journal of Abnormal and Social Psychology,* 1954, *49,* 235–240.

Myers, I. B. *The Myers-Briggs Type Indicator: Manual.* Princeton, N.J.: Educational Testing Service, 1962.

Naylor, J. C., & Shine, L. C. A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology,* 1965, *3,* 33–42.

Nesselroade, J. R., & Reese, H. W. (Eds.) *Life-span developmental psychology: Methodological issues.* New York: Academic Press, 1973.

NEWMAN, S. H. Psychologist, heal thyself! *Psychological Reports*, 1966, *19*, 175–179.

NICHOLS, P. L., & BROMAN, S. H. Familial resemblance in infant mental development. *Developmental Psychology*, 1974, *10*, 442–446.

NISBET, J. D. Symposium: Contributions to intelligence testing and the theory of intelligence: IV. Intelligence and age: Retesting with twenty-four years' interval. *British Journal of Educational Psychology*, 1957, *27*, 190–198.

NORMAN, R. D. A revised deterioration formula for the Wechsler Adult Intelligence Scale. *Journal of Clinical Psychology*, 1966, *22*, 287–294.

NOVICK, M. R., & LEWIS, C. Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, *32*, 1–13.

NOWICKI, S., JR., & DUKE, M. P. A preschool and primary internal-external control scale. *Developmental Psychology*, 1974, *10*, 874–880.

NUNNALLY, J. C. *Educational measurement and evaluation.* (2nd ed.) New York: McGraw-Hill, 1972.

O'CONNOR, E. J., WEXLEY. K. N., & ALEXANDER, R. A. Single-group validity: Fact or fallacy? *Journal of Applied Psychology*, 1975, *60*, 352–355.

OEHRN, A. *Experimentelle Studien zur Individualpsychologie.* Dorpater disser., 1889. (Also in *Psychologische Arbeiten*, 1895, *1*, 95–152.)

OFFICE OF FEDERAL CONTRACT COMPLIANCE (OFCC). Guidelines for reporting criterion-related and content validity. *Federal Register*, 1974, 39(12), 2094–2096.

ORTAR, G. Is a verbal test cross-cultural? *Scripta Hierosolymitana* (Hebrew University, Jerusalem), 1963, *13*, 219–235.

ORTAR, G. Some principles for adaptation of psychological tests. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972. Pp. 111–120.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning.* Urbana, Ill.: University of Illinois Press, 1957.

OSS ASSESSMENT STAFF. *Assessment of men: Selection of personnel for the Office of Strategic Services.* New York: Rinehart, 1948.

OWENS, W. A. Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs*, 1953, *48*, 3–54.

OWENS, W. A. Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology*, 1966, *57*, 311–325.

OWENS, W. A. Toward one discipline of scientific psychology. *American Psychologist*, 1968, *23*, 782–785.

OWENS, W. A., & HENRY, E. R. *Biographical data in industrial psychology: A review and evaluation.* Greensboro, N.C.: Richardson Foundation, 1966.

PALERMO, D. S., & JENKINS, J. J. Frequency of superordinate responses to a word association test as a function of age. *Journal of Verbal Learning and Verbal Behavior*, 1963, *1*, 378–383.

PALMER, J. O. *The psychological assessment of children.* New York: Wiley, 1970.

PALMORE, E. (Ed.) *Normal aging.* Durham, N.C.: Duke University Press, 1970.

PARASKEVOPOULOS, J., & HUNT, J. McV. Object construction and imitation under differing conditions of rearing. *Journal of Genetic Psychology*, 1971, *119*, 301–321.

Pasamanick, B., & Knobloch, H. Retrospective studies on the epidemiology of reproductive casualty: Old and new. *Merrill-Palmer Quarterly*, 1966, *12*, 7–26.

Pascal, G. R., & Suttell, B. J. *The Bender-Gestalt Test: Quantification and validity for adults.* New York: Grune & Stratton, 1951.

Passini, F. T., & Norman, W. T. A universal conception of personality structure. *Journal of Personality and Social Psychology*, 1966, *4*, 44–49.

Patterson, C. H. Predicting success in trade and vocational courses: Review of the literature. *Educational and Psychological Measurement*, 1956, *16*, 352–400.

Patterson, C. H. Interest tests and the emotionally disturbed client. *Educational and Psychological Measurement*, 1957, *17*, 264–280.

Paul, G. L., & Eriksen, C. W. Effects of test anxiety on "real-life" examinations. *Journal of Personality*, 1964, *32*, 480–494.

Pearson, J. S., & Swenson, W. M. *A user's guide to the Mayo Clinic automated MMPI program.* New York: Psychological Corporation, 1967.

Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 1901, *2*, 559–572.

Peel, E. A. A note on practice effects in intelligence tests. *British Journal of Educational Psychology*, 1951, *21*, 122–125.

Peel, E. A. Practice effects between three consecutive tests of intelligence. *British Journal of Educational Psychology*, 1952, *22*, 196–199.

Peizer, E. Studies of the upper extremity amputee. I. Design and scope. *Artificial Limbs*, 1958, *5*(1), 4–56.

Pemberton, C. L. The closure factors related to temperament. *Journal of Personality*, 1952, *21*, 159–175.

Pepinsky, H. B., Siegel, L., & Vanatta, A. The criterion in counseling: A group participation scale. *Journal of Abnormal and Social Psychology*, 1952, *47*, 415–419.

Petersen, N. S. An expected utility model for "optimal" selection. American College Testing Program, Technical Bulletin No. 24, 1974.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*, 3–29.

Peterson, D. R. *The clinical study of social behavior.* New York: Appleton-Century-Crofts, 1968.

Peterson, J. *Early conceptions and tests of intelligence.* Yonkers, N.Y.: World Book Co., 1926.

Philippe, J. Jastrow—exposition d'anthropologie de Chicago—testes psychologiques, etc. *Année psychologique*, 1894, *1*, 522–526.

Pike, L. W., & Evans, F. R. Effects of special instruction for three kinds of mathematics aptitude items. *College Entrance Examination Board Research Report*, 1, 1972.

Pinard, A., & Laurendeau, M. A scale of mental development based on the theory of Piaget: Description of a project. *Journal of Research in Science Teaching*, 1964, *2*, 253–260.

Pinder, C. C. Statistical accuracy and practical utility in the use of moderator variables. *Journal of Applied Psychology*, 1973, *57*, 214–221.

PINNEAU, S. R. *Changes in intelligence quotient from infancy to maturity.* Boston: Houghton Mifflin, 1961.

PIOTROWSKI, Z. A. Digital-computer interpretation of inkblot test data. *Psychiatric Quarterly,* 1964, *38,* 1–26.

PITCHER, B. The Law School Admission Test as a predictor of first-year law school grades, 1962–1963. *Statistical Report, Educational Testing Service,* SR-65–32, 1965.

PITTLUCK, P. The relation between aggressive fantasy and overt behavior. Unpublished doctoral dissertation, Yale University, 1950.

PLANT, W. T., & MINIUM, E. W. Differential personality development in young adults of markedly different aptitude levels. *Journal of Educational Psychology,* 1967, *58,* 141–152.

POPHAM, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), *Criterion-referenced measurement.* Englewood Cliffs, N.J.: Educational Technology Publications, 1971. Pp. 79–98.

POPHAM, W. J., & HUSEK, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, *6,* 1–9.

POSTMAN, L., & KEPPEL, G. *Norms of word association.* New York: Academic Press, 1970.

PRIEN, E. P. Dynamic character of criteria: Organization change. *Journal of Applied Psychology,* 1966, *50,* 501–504.

PRIMOFF, E. S. Empirical validations of the J-coefficient. *Personnel Psychology,* 1959, *12,* 413–418.

PRIMOFF, E. S. *Job element methods. Volume 3: The J-coefficient.* Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, 1975.

Privacy and behavioral research. Washington, D.C.: Government Printing Office, 1967.

Publication manual of the American Psychological Association. Washington, D.C.: American Psychological Association, 1974.

QUERESHI, M. Y. Practice effects on the WISC subtest scores and IQ estimates. *Journal of Clinical Psychology,* 1968, *24,* 79–85.

RABIN, A. I. (Ed.) *Projective techniques in personality assessment.* New York. Springer, 1968.

RABIN, A. I., & GUERTIN, W. H. Research with the Wechsler-Bellevue Test: 1945–1950. *Psychological Bulletin,* 1951, *48,* 211–248.

RADCLIFFE, J. A. A note on questionnaire faking with 16PFQ and MPI. *Australian Journal of Psychology,* 1966, *18,* 154–157.

RAMSEY, P. H., & VANE, J. R. A factor analytic study of the Stanford-Binet with young children. *Journal of School Psychology,* 1970, *8,* 278–284.

RAPAPORT, D., et al. *Diagnostic psychological testing.* Chicago: Year Book Publishers, Vol. I, 1945; Vol. II, 1946.

REES, A. H., & PALMER, F. H. Factors related to change in mental test performance. *Developmental Psychology Monographs,* 1970, *3*(2, Part 2).

REICHENBERG-HACKETT, W. Changes in Goodenough drawings after a gratifying experience. *American Journal of Orthopsychiatry,* 1953, *23,* 501–517.

REID, J. W. Stability of measured Kuder interests in young adults. *Journal of Educational Research,* 1951, *45,* 307–312.

REILLY, R. R. A note on minority group test bias studies. *Psychological Bulletin,* 1973, *80,* 130–132.

REITAN, R. M. Certain differential effects of left and right cerebral lesions in human adults. *Journal of Comparative and Physiological Psychology,* 1955, *48,* 474–477.

REITAN, R. M. A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), *International Review of Research in Mental Retardation,* Vol. 1. New York: Academic Press, 1966. Pp. 153–218.

The responsible use of tests: A position paper of AMEG, APGA, and NCME. *Measurement and Evaluation in Guidance,* 1972, *5,* 385–388.

REYNOLDS, H. H. Efficacy of sociometric ratings in predicting leadership success. *Psychological Reports,* 1966, *19,* 35–40.

RICH, C. C., & ANDERSON, R. P. A tactual form of the Progressive Matrices for use with blind children. *Personnel and Guidance Journal,* 1965, *43,* 912–919.

RICHARDS, J. M., JR., TAYLOR, C. W., PRICE, P. B., & JACOBSEN, T. L. An investigation of the criterion problem for one group of medical specialists. *Journal of Applied Psychology,* 1965, *49,* 79–90.

RICHARDSON, M. W. The interpretation of a test validity coefficient in terms of increased efficiency of a selected group of personnel. *Psychometrika,* 1944, *9,* 245–248.

RIMOLDI, H. J. A note on Raven's Progressive Matrices test. *Educational and Psychological Measurement,* 1948, *8,* 347–352.

ROBACK, H. B. Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. *Psychological Bulletin,* 1968, *70,* 1–19.

ROBB, G. P., BERNARDONI, L. C., & JOHNSON, R. W. *Assessment of individual mental ability.* Scranton: International Textbook Company, 1972.

ROBINSON, D. D. Prediction of clerical turnover in banks by means of a weighted application blank. *Journal of Applied Psychology,* 1972, *56,* 282.

ROBINSON, J. P., & SHAVER, P. R. *Measures of social psychological attitudes.* Ann Arbor, Mich.: Institute for Social Research, The University of Michigan, 1973.

RODGER, A. G. The application of six group intelligence tests to the same children, and the effects of practice. *British Journal of Educational Psychology,* 1936, *6,* 291–305.

ROE, A. *The psychology of occupations.* New York: Wiley, 1956.

ROGERS, C. R., & DYMOND, R. F. (Eds.) *Psychotherapy and personality change.* Chicago: University of Chicago Press, 1954.

RORER, L. G. The great response-style myth. *Psychological Bulletin,* 1965, *63,* 129–156.

RORER, L. G., HOFFMAN, P. J., LA FORGE, G. E., & HSIEH, KUO-CHENG. Optimum cutting scores to discriminate groups of unequal size and variance. *Journal of Applied Psychology,* 1966, *50,* 153–164.

RORSCHACH, H. (Transl. by P. Lemkau & B. Kronenburg.) *Psychodiagnostics: A diagnostic test based on perception.* Berne: Huber, 1942. (1st German ed., 1921; U.S. distr., Grune & Stratton.)

ROSEN, E., & RIZZO, G. B. Preliminary standardization of the MMPI for use in Italy: A case study in intercultural and intracultural differences. *Educational and Psychological Measurement,* 1961, *21,* 629–636.

ROSENBERG, L. A., & STROUD, M. Limitations of brief intelligence testing with young children. *Psychological Reports,* 1966, *19,* 721–722.

ROSENBERG, N. Stability and maturation of Kuder interest patterns of medical, law, and business school alumni. *Journal of Applied Psychology,* 1953, *37,* 367–369.

ROSENTHAL, R. *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts, 1966.

ROSENTHAL, R., & ROSNOW, R. L. (Eds.) *Artifact in behavioral research.* New York: Academic Press, 1969.

ROTTER, J. B. *Social learning and clinical psychology.* Englewood Cliffs, N.J.: Prentice-Hall, 1954.

ROTTER, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 1966, *80*(1, Whole No. 609).

RUCH, W. W. A re-analysis of published differential validity studies. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972.

RUEBHAUSEN, O. M., & BRIM, O. G., JR. Privacy and behavioral research. *American Psychologist,* 1966, *21,* 423–437.

RULON, P. J. A simplified procedure for determining the reliability of a test of split-halves. *Harvard Educational Review,* 1939, 9, 99–103.

RUSSELL SAGE FOUNDATION. *Guidelines for the collection, maintenance, and dissemination of pupil records.* New York: Author, 1970.

SACHMAN, H. An investigation of certain aspects of the validity of the formal Rorschach scoring system in relation to age, education, and vocabulary score. Unpublished doctoral dissertation, Fordham University, 1952.

SACKS, E. L. Intelligence scores as a function of experimentally established social relationships between the child and examiner. *Journal of Abnormal and Social Psychology,* 1952, *47,* 354–358.

SAMELSON, F. Response style: A psychologist's fallacy? *Psychological Bulletin,* 1972, *78,* 13–16.

SANFORD, N. (Ed.) Personality development during the college years. *Journal of Social Issues,* 1956, *12,* 3–70.

SARASON, I. G. Test anxiety and the intellectual performance of college students. *Journal of Educational Psychology,* 1961, *52,* 201–206.

SARASON, I. G. *Personality: An objective approach.* (2nd ed.) New York: Wiley, 1972.

SARASON, S. B. *The clinical interaction, with special reference to the Rorschach.* New York: Harper, 1954.

SARASON, S. B., DAVIDSON, K. S., LIGHTHALL, F. F., WAITE, R. R., & RUEBUSH, B. K. *Anxiety in elementary school children.* New York: Wiley, 1960.

SARASON, S. B., & GLADWIN, T. *Psychological problems in mental deficiency.* (3rd ed.) New York: Harper, 1959.

SARASON, S. B., HILL, K. T., & ZIMBARDO, P. A longitudinal study of the relation of test anxiety to performance on intelligence and achievement tests. *Monographs of the Society for Research in Child Development,* 1964, *29*(7, Serial No. 98).

SARBIN, T. R., TAFT, R., & BAILEY, D. E. *Clinical inference and cognitive theory.* New York: Holt, Rinehart and Winston, 1960.

SATTLER, J. M. Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. *Psychological Bulletin,* 1970, *73,* 137–160.

SATTLER, J. M. *Assessment of children's intelligence.* Philadelphia: Saunders, 1974 (1975 revised reprint).

SATTLER, J. M., & THEYE, F. Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin,* 1967, *68,* 347–360.

SAUNDERS, D. R. Moderator variables in prediction. *Educational and Psychological Measurement,* 1956, *16,* 209–222.

SAWYER, J. Measurement *and* prediction, clinical *and* statistical. *Psychological Bulletin,* 1966, *66,* 178–200.

SCHAIE, K. W. A general model for the study of developmental problems. *Psychological Bulletin,* 1965, *64,* 92–107.

SCHAIE, K. W., & LABOUVIE-VIEF, G. Generational versus ontogenetic components of change in cognitive behavior: A fourteen-year cross-sequential study. *Developmental Psychology,* 1974, *10,* 305–320.

SCHAIE, K. W., LABOUVIE, G. V., & BUECH, B. U. Generational and cohort-specific differences in adult cognitive functioning: A fourteen-year study of independent samples. *Developmental Psychology,* 1973, 9, 151–166.

SCHAIE, K. W., & STROTHER, C. R. A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin,* 1968, *70,* 671–680.

SCHMIDT, F. L., BERNER, J. G., & HUNTER, J. E. Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology,* 1973, *58,* 5–9.

SCHMIDT, F. L., & HUNTER, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist,* 1974, *29,* 1–8.

SCHUTZ, R. E. SRA Primary Mental Abilities, 1962 Edition. *Seventh Mental Measurements Yearbook,* 1972, 1067.

SCHWARZ, P. A. Development and application of African ability tests: Summary Report. American Institutes for Research in the Behavioral Sciences, Contract ICAc-2155, AIR-C71-12/64-TR, 1964. (a)

SCHWARZ, P. A. Development of manpower screening tests for the developing nations: Technical manual. American Institutes for Research in the Behavioral Sciences, Contract ICAc-2155, AIR-C71-6/64-TR, 1964. (b)

SCHWARZ, P. A., & KRUG, R. E. *Ability testing in developing countries: A handbook of principles and techniques.* New York: Praeger, 1972.

SCOTT, W. A. Comparative validities of forced-choice and single-stimulus tests. *Psychological Bulletin,* 1968, *70,* 231–244.

SCOTT, W. E., JR., & DAY, G. J. Personality dimensions and vocational interests among graduate business students. *Journal of Counseling Psychology,* 1972, *19,* 30–36.

SEASHORE, C. E. *Psychology of music.* New York: McGraw-Hill, 1938.

SEASHORE, H. G. Women are more predictable than men. *Journal of Counseling Psychology,* 1962, *9,* 261–270.

SEASHORE, H. G., WESMAN, A. G., & DOPPELT, J. E. The standardization of the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology,* 1950, *14,* 99–110.

SEASHORE, S. E., INDIK, B. P., & GEORGOPOULOS, B. S. Relationships among criteria of job performance. *Journal of Applied Psychology,* 1960, *44,* 195–202.

SECHREST, L. Incremental validity: A recommendation. *Educational and Psychological Measurement,* 1963, *23,* 153–158.

SEGALL, M. H., CAMPBELL, D. T., & HERSKOVITS, M. J. *The influence of culture on visual perception.* Indianapolis: Bobbs-Merrill, 1966.

SEGUIN, E. *Idiocy: Its treatment by the physiological method.* (Reprinted from original ed. of 1866.) New York: Bureau of Publications, Teachers College, Columbia University, 1907.

SHARP, S. E. Individual psychology: A study in psychological method. *American Journal of Psychology,* 1898–99, *10,* 329–391.

SHAW, M. E., & WRIGHT, J. M. *Scales for the measurement of attitudes.* New York: McGraw-Hill, 1967.

SHERIF, C. W., SHERIF, M., & NEBERGALL, R. *Attitude and attitude change: The social judgment-involvement approach.* Philadelphia: Saunders, 1965.

SHIMBERG, M. E. An investigation into the validity of norms with special reference to urban and rural groups. *Archives of Psychology,* 1929, No. 104.

SHIPLEY, W. C., GRAY, F. E., & NEWBERT, N. The personal inventory. *Journal of Clinical Psychology,* 1946, *2,* 318–322.

SHNEIDMAN, E. S. Projective techniques. In B. B. Wolman (Ed.), *Handbook of clinical psychology.* New York: McGraw-Hill, 1965. Pp. 498–521.

SHORE, C. W., & MARION, R. Suitability of using common selection test standards for Negro and white airmen. Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. AFHRL-TR-72–53, May 1972.

SIESS, T. F., & JACKSON, D. N. A personological approach to the interpretation of vocational interests. *Proceedings, 75th Annual Convention, American Psychological Association,* 1967, 353–354.

SIGEL, I. E. How intelligence tests limit understanding of intelligence. *Merrill-Palmer Quarterly,* 1963, 9, 39–56.

SIGEL, I. E. Where is preschool education going: Or are we en route without a map? *Proceedings, 1972 Invitational Conference on Testing Problems, Educational Testing Service,* 1973, 99–116.

SILVERMAN, I., & SHULMAN, A. D. A conceptual model of artifact in attitude change studies. *Sociometry,* 1970, *33,* 97–107.

SILVERMAN, L. H. A Q-sort of the validity of evaluations made from projective techniques. *Psychological Monographs,* 1959, *73*(7, Whole No. 477).

SILVERSTEIN, A. B. Validity of WPPSI short forms. *Journal of Consulting and Clinical Psychology,* 1968, *32,* 229–230. (a)

SILVERSTEIN, A. B. Validity of a new approach to the design of WAIS, WISC, and WPPSI short forms. *Journal of Consulting and Clinical Psychology,* 1968, *32,* 478–479. (b)

SILVERSTEIN, A. B. Reappraisal of the validity of WAIS, WISC, and WPPSI short forms. *Journal of Consulting and Clinical Psychology,* 1970, *34,* 12–14.

SILVERSTEIN, A. B. A corrected formula for assessing the validity of WAIS, WISC, and WPPSI short forms. *Journal of Clinical Psychology,* 1971, *27,* 212–213.

SILVERSTEIN, A. B. Factor structure of the Wechsler Intelligence Scale for Children for three ethnic groups. *Journal of Educational Psychology,* 1973, *65,* 408–410.

SISSON, D. E. Forced-choice—the new Army rating. *Personnel Psychology,* 1948, *1,* 365–381.

SKOLNICK, A. Motivational imagery and behavior over twenty years. *Journal of Consulting Psychology,* 1966, *30,* 463–478.

SLOAN, W. The Lincoln-Oseretsky Motor Development Scale. *Genetic Psychology Monographs,* 1955, *51,* 183–252.

SMITH, A. Talkers and doers: Or education, intelligence, and WAIS verbal-performance ratios in psychiatric patients. *American Psychologist,* 1966, *21,* 687.

SMITH, G. H. *Motivation research in advertising and marketing.* New York: McGraw-Hill, 1954.

SMITH, P. C., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology,* 1963, *47,* 149–155.

SMITH, S. Language and non-verbal test performance of racial groups in Honolulu before and after a 14-year interval. *Journal of General Psychology,* 1942, *26,* 51–93.

SNIDER, J. G., & OSGOOD, C. E. (Eds.) *Semantic differential technique: A sourcebook.* Chicago: Aldine, 1969.

SNYDER, C. R., & LARSON, G. R. A further look at student acceptance of general personality interpretations. *Journal of Consulting and Clinical Psychology,* 1972, *38,* 384–388.

SNYGG, D., & COMBS, A. W. *Individual behavior: A perceptual approach to behavior.* (Rev. ed.) New York: Harper, 1959.

SOMMER, R. *Diagnostik der Geisteskrankheiten für praktische Ärzte und Studierende.* Wien und Leipzig: Urban und Schwarzenberg, 1894.

SONTAG, L. W., BAKER, C. T., & NELSON, V. L. Mental growth and personality development: A longitudinal study. *Monographs of the Society for Research in Child Development,* 1958, *23*(2, Serial No. 68).

SORENSON, W. W. Test of mechanical principles as a suppressor variable for the prediction of effectiveness on a mechanical repair job. *Journal of Applied Psychology,* 1966, *50,* 348–352.

SPEARMAN, C. "General intelligence" objectively determined and measured. *American Journal of Psychology,* 1904, *15,* 201–293.

SPEARMAN, C. *The abilities of man.* New York: Macmillan, 1927.

SPIELBERGER, C. D. (Ed.) *Anxiety: Current trends in theory and research.* Vol. 2. New York: Academic Press, 1972.

SPRANGER, E. (transl. by P. J. W. Pigors). *Types of men.* Halle: Niemeyer, 1928.

STAGNER, R. The gullibility of personnel managers. *Personnel Psychology,* 1958, *11,* 347–352.

Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

STANLEY, J. C. (Ed.) *Preschool programs for the disadvantaged: Five experimental approaches to early childhood education.* Baltimore, Md.: Johns Hopkins University Press, 1972.

STANLEY, J. C. (Ed.) *Compensatory education for children, ages two to eight.* Baltimore, Md.: Johns Hopkins University Press, 1973.

STANLEY, J. C., & HOPKINS, K. D. *Educational and psychological measurement and evaluation.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

STANLEY, J. C., & PORTER, A. C. Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. *Journal of Educational Measurement,* 1967, *4,* 199–218.

STARR, B. J., & KATKIN, E. S. The clinician as aberrant actuary: Illusory correlation and the Incomplete Sentences Blank. *Journal of Abnormal Psychology,* 1969, *74,* 670–675.

STECKLER, V. A review of the Nonreading Aptitude Test Battery. *Journal of Employment Counseling,* 1973, *10,* 17–20.

STEPHENS, M. W., & DELYS, P. A locus of control measure for preschool children. *Developmental Psychology,* 1973, *9,* 55–65.

STEPHENSON, W. *The study of behavior: Q-technique and its methodology.* Chicago: University of Chicago Press, 1953.

STERN, G. G. *People in context: Measuring person-environment congruence in education and industry.* New York: Wiley, 1970.

STOTT, L. H., & BALL, R. S. Infant and preschool mental tests: Review and evaluation. *Monographs of the Society for Research in Child Development.* 1965, *30*(3, Serial No. 101).

STRAUSS, A. A., & LEHTINEN, L. E. *Psychopathology and education of the brain-injured child.* New York: Grune & Stratton, 1947.

STRICKER, L. J. Compulsivity as a moderator variable: A replication and extension. *Journal of Applied Psychology,* 1966, *50,* 331–335.

STRICKER, L. J. "Test-wiseness" on personality scales. *Journal of Applied Psychology Monograph,* 1969, *53*(3, Part 2).

STRONG, D. J., & FEDER, D. D. Measurement of the self-concept: A critique of the literature. *Journal of Counseling Psychology,* 1961, *8,* 170–178.

STRUNK, W., JR., & WHITE, E. B. *The elements of style.* (2nd ed.). New York: Macmillan, 1972.

SUMMERS, G. F. (Ed.) *Attitude measurement.* Chicago: Rand McNally, 1970.

SUNDBERG, N. D. The acceptability of "fake" versus "bona fide" personality test interpretations. *Journal of Abnormal and Social Psychology,* 1955, *50,* 145–147.

SUNDBERG, N. D. The use of the MMPI for cross-cultural personality study: A preliminary report on the German translation. *Journal of Abnormal and Social Psychology,* 1956, *58,* 281–283.

SUPER, D. E. *The psychology of careers: an introduction to vocational development.* New York: Harper & Row, 1957.

SUPER, D. E. (Ed.) *The use of multifactor tests in guidance.* Washington, D.C.: American Personnel and Guidance Association, 1958. (Reprinted from *Personnel and Guidance Journal*, 1956, 1957).

SUPER, D. E. *DAT Career Planning Program: Counselor's manual.* New York: Psychological Corporation, 1973.

SUPER, D. E., et al. *Vocational development: A framework for research.* New York: Teachers College Press, 1957.

SUPER, D. E., et al. *Computer-assisted counseling.* New York: Teachers College Press, 1970.

SUPER, D. E., & CRITES, J. O. *Appraising vocational fitness by means of psychological tests.* (Rev. ed.) New York: Harper, 1962.

SWENSEN, C. H. Empirical evaluations of human figure drawings: 1957–1966. *Psychological Bulletin*, 1968, *70*, 20–44.

SWENSON, W. M., PEARSON, J. S., & OSBORNE, D. *An MMPI source book: Basic item, scale, and pattern data on 50,000 medical patients.* Minneapolis: University of Minnesota Press, 1973.

SWETS, J. A., & ELLIOTT, L. L. (Eds.) *Psychology and the handicapped child.* Washington, D.C.: U.S. Government Printing Office, 1974. (DHEW Publication No. OE-73-05000.)

SYMONDS, P. M. *Diagnosing personality and conduct.* New York: Century, 1931.

TAFT, R. A cross-cultural comparison of the MMPI. *Journal of Consulting Psychology*, 1957, *21*, 161–164.

TANOFSKY, R., SHEPPS, R. R., & O'NEILL, P. J. Pattern analysis of biographical predictors of success as an insurance salesman. *Journal of Applied Psychology*, 1969, *53*, 136–139.

TASTO, D. L., HICKSON, R., & RUBIN, S. E. Scaled profile analysis of fear survey schedule factors. *Behavior Therapy*, 1971, *2*, 543–549.

TAYLOR, C. W. (Ed.) *Climate for creativity: Report of the Seventh National Research Conference on Creativity.* New York: Pergamon Press, 1972.

TAYLOR, C. W., & BARRON, F. (Eds.) *Scientific creativity: Its recognition and development.* New York: Wiley, 1963.

TAYLOR, H. C., & RUSSELL, J. T. The relationship of validity coefficients to .the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 1939, *23*, 565–578.

TAYLOR, J. A. A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology*, 1953, *48*, 285–290.

TEAHAN, J. E., & DREWS, E. M. A comparison of northern and southern Negro children on the WISC. *Journal of Consulting Psychology*, 1962, *26*, 292.

TELLEGEN, A., & BRIGGS, P. F. Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, 1967, *31*, 499–506.

TEMP, G. Test bias: Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement*, 1971, *8*, 245–251.

TERMAN, L. M. *The measurement of intelligence.* Boston: Houghton Mifflin, 1916.

TERMAN, L. M., & MERRILL, M. A. *Measuring intelligence.* Boston: Houghton Mifflin, 1937.

TERMAN, L. M., & MERRILL, M. A. *Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M.* Boston: Houghton Mifflin, 1960.

TERMAN, L. M., & MERRILL, M. A. *Stanford-Binet Intelligence Scale: 1972 norms edition.* Boston: Houghton Mifflin, 1973.

TERMAN, L. M., & ODEN, M. H. *The gifted child grows up: Twenty-five years' follow-up of a superior group.* Stanford, Calif.: Stanford University Press, 1947.

TERMAN, L. M., & ODEN, M. H. *The gifted group at mid-life: Thirty-five years' follow-up of the superior child.* Stanford, Calif.: Stanford University Press, 1959.

Testing and public policy. (Special issue) *American Psychologist,* 1965, *20,* 857–992.

THARP, R. G., & WETZEL, R. J. *Behavior modification in the natural environment.* New York: Academic Press, 1969.

THEOLOGUS, G. C., & FLEISHMAN, E. A. *Development of a taxonomy of human performance; Validation study of ability scales for classifying human tasks.* Washington, D.C.: American Institutes for Research, 1971.

THOMAS, H. Psychological assessment instruments for use with human infants. *Merrill-Palmer Quarterly,* 1970, *16,* 179–223.

THOMSON, G. H. A hierarchy without a general factor. *British Journal of Psychology,* 1916, *8,* 271–281.

THOMSON, G. H. *The factorial analysis of human ability.* (3rd ed.) Boston: Houghton Mifflin, 1948.

THORNDIKE, R. L. The effect of interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology,* 1933, *24,* 543–549.

THORNDIKE, R. L. "Constancy" of the IQ. *Psychological Bulletin,* 1940, *37,* 167–186.

THORNDIKE, R. L. *Personnel selection: Test and measurement techniques.* New York: Wiley, 1949.

THORNDIKE, R. L. *The concepts of over- and under-achievement.* New York: Teachers College Press, 1963.

THORNDIKE, R. L. Concepts of culture-fairness. *Journal of Educational Measurement,* 1971, *8,* 63–70.

THURSTONE, L. L. *Vectors of mind: Multiple-factor analysis for the isolation of primary traits.* Chicago: University of Chicago Press, 1935.

THURSTONE, L. L. Primary mental abilities. *Psychometric Monographs,* 1938, No. 1.

THURSTONE, L. L. A factorial study of perception. *Psychometric Monographs,* 1944, No. 4.

THURSTONE, L. L. *Multiple factor analysis.* Chicago: University of Chicago Press, 1947.

THURSTONE, L. L. Some primary abilities in visual thinking. *Psychometric Laboratory, University of Chicago,* No. 59, August 1950.

THURSTONE, L. L. Creative talent. *Proceedings of the 1950 Invitational Conference on Testing Problems, Educational Testing Service,* 1951, 55–69. (Reprinted in A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 414–428.)

THURSTONE, L. L. *The measurement of values.* Chicago: University of Chicago Press, 1959.

THURSTONE, L. L., & CHAVE, E. J. *The measurement of attitude.* Chicago: University of Chicago Press, 1929.

THURSTONE, L. L., & THURSTONE, T. G. Factorial studies of intelligence. *Psychometric Monographs,* 1941, No. 2.

TIEBOUT, C., & MEIER, N. C. Artistic ability and general intelligence. *Psychological Monographs,* 1936, *48*(1, Whole No. 213).

TOLOR, A., & SCHULBERG, H. C. *An evaluation of the Bender-Gestalt Test.* Springfield, Ill.: Charles C Thomas, 1963.

TORRANCE, E. P. *Guiding creative talent.* Englewood Cliffs, N.J.: Prentice-Hall, 1962.

TORRANCE, E. P. *Education and the creative potential.* Minneapolis: University of Minnesota Press, 1963.

TORRANCE, E. P. *Rewarding creative behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1965.

TRACHT, V. S. Preliminary findings on testing the cerebral palsied with Raven's "Progressive Matrices." *Journal of Exceptional Children,* 1948, *15,* 77–79.

TRAXLER, A. E., & HILKERT, R. N. Effect of type of desk on results of machine-scored tests. *School and Society,* 1942, *56,* 277–296.

TRYON, R. C. A theory of *psychological components*—an alternative to "mathematical factors." *Psychological Review,* 1935, *42,* 425–454.

TSUDZUKI, A., HATA, Y., & KUZE, T. (A study of rapport between examiner and subject.) *Japanese Journal of Psychology,* 1957, *27,* 22–28.

TUCKER, L. R. Experiments in multimode factor analysis. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 369–379.

TUDDENHAM, R. D. Soldier intelligence in World Wars I and II. *American Psychologist,* 1948, *3,* 54–56.

TUDDENHAM, R. D., BLUMENKRANTZ, J., & WILKIN, W. R. Age changes on AGCT: A longitudinal study of average adults. *Journal of Counseling and Clinical Psychology,* 1968, *32,* 659–663.

TYLER, L. E. Toward a workable psychology of individuality. *American Psychologist,* 1959, *14,* 75–81.

TYLER, L. E. A significant change in direction: Implications for measurement. *Proceedings, 1973 Invitational Conference on Testing Problems, Educational Testing Service,* 1974, 70–78.

TYLER, L. E., SUNDBERG, N. D., ROHILA, P. K., & GREENE, M. M. Patterns of choices in Dutch, American, and Indian adolescents. *Journal of Counseling Psychology,* 1968, *15,* 522–529.

UÇMAN, P. A normative study of the Goodenough-Harris Drawing Test on a Turkish sample. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972. Pp. 365–370.

UHRBROCK, R. S. Attitudes of 4430 employees. *Journal of Social Psychology,* 1934, *5,* 365–377.

ULRICH, L., & TRUMBO, D. The selection interview since 1949. *Psychological Bulletin,* 1965, *63,* 100–116.

ULRICH, R., STACHNIK, T., & STAINTON, R. Student acceptance of generalized personality interpretations. *Psychological Reports,* 1963, *13,* 831–834.

UPSHAW, H. S. The effects of variable perspectives on judgments of opinion statements for Thurstone scales: Equal-appearing intervals. *Journal of Personality and Social Psychology,* 1965, *2,* 60–69.

UNITED STATES DEPARTMENT OF LABOR: MANPOWER ADMINISTRATION. Technical report on standardization of the General Aptitude Test Battery. *Technical Report* S-121, July 1958.

UNITED STATES DEPARTMENT OF LABOR: MANPOWER ADMINISTRATION. *Manual for the USES General Aptitude Test Battery.* Washington, D.C.: U.S. Employment Service, 1970. (a)

UNITED STATES DEPARTMENT OF LABOR: MANPOWER ADMINISTRATION. *Manual for the USES Nonreading Aptitude Test Battery.* Washington, D.C. U.S. Employment Service, 1970. (b)

UZGIRIS, I. Č., & HUNT, J. McV. *Assessment in infancy: Ordinal Scales of Psychological Development.* Urbana, Ill.: University of Illinois Press, 1975.

VAN DER MADE-VAN BEKKUM, I. J. *Dutch word association norms.* Amsterdam: Swets & Zeitlinger N. V., 1971.

VARBLE, D. L. Current status of the Thematic Apperception Test. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 2. Palo Alto, Calif.: Science and Behavior Books, 1971. Pp. 216–235.

VELICER, W. F. Comment on the general inapplicability of Ghiselli's moderator system for two predictors. *Journal of Applied Psychology,* 1972, *56,* 262–265. (a)

VELICER, W. F. The moderator variable viewed as heterogeneous regression. *Journal of Applied Psychology,* 1972, *56,* 266–269. (b)

VERNON, P. E. The validation of civil service selection board procedures. *Occupational Psychology,* 1950, *24,* 75–95.

VERNON, P. E. *The structure of human abilities.* (Rev. ed.) London: Methuen, 1960.

VERNON, P. E. *Intelligence and cultural environment.* London: Methuen, 1969.

VEROFF, J., ATKINSON, J. W., FELD, S. C., & GURIN, G. The use of thematic apperception to assess motivation in a nationwide interview study. In J. W. Atkinson, J. O. Raynor, et al. *Motivation and achievement.* Washington, D.C.: Winston (distr. by Halsted Press Division, Wiley), 1974. Ch. 3.

VINITSKY, M. A forty-year follow-up on the vocational interests of psychologists and their relationship to career development. *American Psychologist,* 1973, *28,* 1000–1009.

VITOLA, B. M., & ALLEY, W. E. Development and standardization of Air Force Composites for the Armed Services Vocational Aptitude Battery. Lackland AFB, Texas: Air Forces Human Resources Laboratory. AFHRL-TR-68–110, September 1968.

WAGGENER, R., STARR, R. P., & HOPKINS, T. Norming of a basic learning examination and its use to predict GED results. Paper presented at the meeting of the American Personnel and Guidance Association, New Orleans, April 1974.

WAHLSTROM, M., & BOERSMAN, F. J. The influence of test-wiseness upon achievement. *Educational and Psychological Measurement,* 1968, *28,* 413–420.

WAINWRIGHT, W. H. Cultural attitudes and clinical judgment. *International Journal of Social Psychiatry,* 1958, *4,* 105–107.

WAITE, R. R., SARASON, S. B., LIGHTHALL, F. F., & DAVIDSON, K. S. A study of anxiety and learning in children. *Journal of Abnormal and Social Psychology,* 1958, *57,* 267–270.

WALBERG, H. J., & MARJORIBANKS, K. Differential mental abilities and home environment: A canonical analysis. *Developmental Psychology,* 1973, 9, 363–368.

WALD, A. *Sequential analysis.* New York: Wiley, 1947.

WALD, A. *Statistical decision function.* New York: Wiley, 1950.

WALKER, D. K. *Socioemotional measures for preschool and kindergarten children.* San Francisco: Jossey-Bass, 1973.

WALLACE, S. R. Criteria for what? *American Psychologist,* 1965, *20,* 411–417.

WALLACE, W. L. College Board Scholastic Aptitude Test. *Seventh Mental Measurements Yearbook,* 1972, 648–650.

WALLACH, M. A., & WING, C. W., JR. *The talented student: A validation of the creativity-intelligence distinction.* New York: Holt, Rinehart and Winston, 1969.

WALSH, T. M. Responses on the Famous Sayings Test of professional and nonprofessional personnel in a medical population. *Psychological Reports,* 1966, *18,* 151–157.

WALSH, W. B. Theories of person-environment interaction: Implications for the college student. *American College Testing Program, Monograph No.* 10, 1973.

WARNER, W. L., MEEKER, M., & EELLS, K. *Social class in America: A manual of procedure for the measurement of social status.* Chicago: Science Research Associates, 1949.

WASIK, B. H., & WASIK, J. L. Performance of culturally deprived children on the Concept Assessment Kit—Conservation. *Child Development,* 1971, *42,* 1586–1590.

WATSON, C. G., FELLING, J., & MACEACHERN, D. G. Objective Draw-a-Person scales: An attempted cross-validation. *Journal of Clinical Psychology,* 1967, *23,* 382–386.

WATSON, J. B., & WATSON, R. R. Studies in infant psychology. *Scientific Monthly,* 1921, *13,* 493–515.

WAUGH, K. W., & BUSH, W. J. *Diagnosing learning disorders.* Columbus, Ohio: Chas. E. Merrill, 1971.

WEBB, E. J., CAMPBELL, D. T., SCHWARTZ, R. D., & SECHREST, L. *Unobtrusive measures: Non-reactive research in the social sciences.* Chicago: Rand McNally, 1966.

WECHSLER, D. *The measurement of adult intelligence.* Baltimore: Williams & Wilkins, 1939.

WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corporation, 1955.

WECHSLER, D. *The measurement and appraisal of adult intelligence.* (4th ed.) Baltimore: Williams & Wilkins, 1958.

WECHSLER, D. *Manual: Wechsler Intelligence Scale for Children—Revised.* New York: Psychological Corporation, 1974.

WEISS, D. J. Strategies of adaptive ability measurement. Psychometric Methods Program, Department of Psychology, University of Minnesota, Research Report 74–5, December 1974.

WEISS, D. J., & BETZ, N. E. Ability measurement: Conventional or adaptive. Psychometric Methods Program, Department of Psychology, University of Minnesota, Research Report 73–1, February 1973.

WEISSENBERG, P., & GRUENFELD, L. W. Relationships among leadership dimensions and cognitive style. *Journal of Applied Psychology,* 1966, *50,* 392–395.

WEITZ, J. Selecting supervisors with peer ratings. *Personnel Psychology,* 1958, *11,* 25–35.

WELSH, G. S. Adjective Check List descriptions of Freud and Jung. *Journal of Personality Assessment,* 1975, *39,* 160–168.

WENDER, P. H. *Minimal brain dysfunction in children.* New York: Wiley-Interscience, 1971.

WERNER, E. E., HONZIK, M. P., & SMITH, R. S. Prediction of intelligence and achievement at ten years from twenty months pediatric and psychologic examinations. *Child Development,* 1968, *39,* 1063–1075.

WERNER, H., & STRAUSS, A. A. Pathology of figure-background relation in the child. *Journal of Abnormal and Social Psychology,* 1941, *36,* 236–248.

WERNER, H., & STRAUSS, A. A. Impairment in thought processes of brain-injured children. *American Journal of Mental Deficiency,* 1943, *47,* 291–295.

WESCHLER, I. R., & BERNBERG, R. Indirect methods of attitude measurement. *International Journal of Opinion and Attitude Research,* 1950, *4,* 209–228.

WESMAN, A. G. Effect of speed on item-test correlation coefficients. *Educational and Psychological Measurement,* 1949, *9,* 51–57.

WESMAN, A. G. Faking personality test scores in a simulated employment situation. *Journal of Applied Psychology,* 1952, *36,* 112–113.

WESMAN, A. G. Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971. Ch. 4.

WHEELER, L. R. A comparative study of the intelligence of East Tennessee mountain children. *Journal of Educational Psychology,* 1942, *33,* 321–334.

WHIMBEY, A. E. *Intelligence can be taught.* New York: Dutton, 1975.

WHIMBEY, A. E., & DENENBERG, V. H. Programming life histories: Creating individual differences by the experimental control of early experiences. *Multivariate Behavioral Research,* 1966, *1,* 279–286.

WHITEMAN, M. Intelligence and learning. *Merrill-Palmer Quarterly,* 1964, *10,* 297–309.

WICKES, T. A., JR. Examiner influence in a testing situation. *Journal of Consulting Psychology,* 1956, *20,* 23–26.

WIENER, G., RIDER, R. V., & OPPEL, W. Some correlates of IQ change in children. *Child Development,* 1963, *34,* 61–67.

WIGGINS, J. S. Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin,* 1962, *59,* 224–242.

WIGGINS, J. S. Social desirability estimation and "faking good" well. *Educational and Psychological Measurement,* 1966, *26,* 329–341. (a)

WIGGINS, J. S. Substantive dimensions of self-report in the MMPI item pool. *Psychological Monographs,* 1966, *80*(22, Whole No. 630). (b)

WIGGINS, J. S. *Personality and prediction: Principles of personality assessment.* Reading, Mass.: Addison-Wesley, 1973.

WIGGINS, J. S., & WINDER, C. L. The Peer Nomination Inventory: An empirically derived sociometric measure of adjustment in preadolescent boys. *Psychological Reports,* 1961, *9,* 643–677. (a)

WIGGINS, J. S., & WINDER, C. L. *Peer Nomination Inventory: Technical supplement.* Stanford, Calif.: Stanford University Press, 1961. (b)

WIGGINS, N. Individual viewpoints of social desirability. *Psychological Bulletin,* 1966, *66,* 68–77.

WIGGINS, N., & HOFFMAN, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology,* 1968, *73,* 70–77.

WILLARD, L. S. A comparison of Culture Fair Test scores with group and individual intelligence test scores of disadvantaged Negro children. *Journal of Learning Disabilities,* 1968, *1,* 584–589.

WILLIAMS, M. The effect of past experience on mental performance in the elderly. *British Journal of Medical Psychology,* 1960, *33,* 215–219.

WILLINGHAM, W. W. Predicting success in graduate education. *Science,* 1974, *183,* 273–278.

WILLIS, J. Group versus individual intelligence tests in one sample of emotionally disturbed children. *Psychological Reports,* 1970, *27,* 819–822.

WING, H. D. A factorial study of musical tests. *British Journal of Psychology,* 1941, *31,* 341–355.

WING, H. D. A revision of the Wing Musical Aptitude Test. *Journal of Research in Music Education,* 1962, *10,* 39–46.

WISSLER, C. The correlation of mental and physical traits. *Psychological Monographs,* 1901, *3*(6, Whole No. 16).

WITKIN, H., et al. Social conformity and psychological differentiation. *International Journal of Psychology,* 1974, *9,* 11–29.

WITKIN, H. A., DYK, R. B., FATERSON, H. F., GOODENOUGH, D. R., & KARP, S. A. *Psychological differentiation: Studies of development.* Potomac, Md.: Erlbaum, 1974 (distr. by Halsted Press Division, Wiley).

WITKIN, H. A., LEWIS, H. B., HERTZMAN, M., MACHOVER, K., MEISSNER, P. B., & WAPNER, S. *Personality through perception: An experimental and clinical study.* New York: Harper, 1954. (Reprinted Westport, Conn.: Greenwood, 1972)

WOHLWILL, J. F. Methodology and research strategy in the study of developmental change. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology: Research and theory.* New York: Academic Press, 1970. Pp. 149–191.

WOLF, R. The measurement of environments. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D.C.: American Council on Education, 1966. Pp. 491–503.

WOLF, T. H. *Alfred Binet.* Chicago: University of Chicago Press, 1973.

WOLK, R. L., & WOLK, R. B. *Manual: Gerontological Apperception Test.* New York: Behavioral Publications, 1971.

WOLPE, J. *The practice of behavior therapy.* (2nd ed.) New York: Pergamon, 1973.

WOLPE, J., & LANG, P. J. A fear survey schedule for use in behavior therapy. *Behaviour Research and Therapy,* 1964, *2,* 27–30.

WOMER, F. B. *What is National Assessment?* Ann Arbor, Mich.: National Assessment of Educational Progress, 1970.

WRIGHT, H. F. Observational child study. In P. H. Mussen (Ed.), *Handbook of research methods in child development.* New York: Wiley, 1960. Ch. 3.

WRIGHTSTONE, J. W., HOGAN, T. P., & ABBOTT, M. M. Accountability in education and associated measurement problems. *Test Service Notebook 33.* New York: Harcourt Brace Jovanovich, 1972.

YAMAMOTO, K., & FRENGEL, B. A. An exploratory component analysis of the Minnesota tests of creative thinking. *California Journal of Educational Research,* 1966, *17,* 220–229.

YATES, A. J., et al. Symposium on the effects of coaching and practice in intelligence tests. *British Journal of Educational Psychology,* 1953, *23,* 147–162; 1954, *24,* 1–8, 57–63.

YERKES, R. M. (Ed.) Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences,* 1921, *15.*

ZEDECK, S., IMPARATO, N., KRAUSZ, M., & OLENO, T. Development of behaviorally anchored rating scales as a function of organizational level. *Journal of Applied Psychology,* 1974, *59,* 249–252.

ZIMMERMAN, B. J., & ROSENTHAL, T. L. Conserving and retaining equalities and inequalities through observation and correction. *Developmental Psychology,* 1974, *10,* 260–268. (a)

ZIMMERMAN, B. J., & ROSENTHAL, T. L. Observational learning of rule-governed behavior by children. *Psychological Bulletin,* 1974, *81,* 29–42. (b)

ZIMMERMAN, I. L., & WOO-SAM, J. M. Research with the Wechsler Intelligence Scale for Children: 1960–1970. *Psychology in the Schools,* 1972, *9,* 232–271. (Special Monograph Supplement)

ZIMMERMAN, I. L., WOO-SAM, J. M., & GLASSER, A. J. *The clinical interpretation of the Wechsler Adult Intelligence Scale.* New York: Grune & Stratton, 1973.

ZUBIN, J., ERON, L. D., & SCHUMER, F. *An experimental approach to projective techniques.* New York: Wiley, 1965.

APPENDIXES

A. Ethical Standards of Psychologists

B. Guidelines on Employee Selection Procedures (EEOC)

Guidelines for Reporting Criterion-Related and Content Validity (OFCC)

C. A Suggested Outline for Test Evaluation

D. Test Publishers

E. Classified List of Representative Tests

APPENDIX A

Ethical Standards of Psychologists[1]

The psychologist believes in the dignity and worth of the individual human being. He is committed to increasing man's understanding of himself and others. While pursuing this endeavor, he protects the welfare of any person who may seek his service or of any subject, human or animal, that may be the object of his study. He does not use his professional position or relationships, nor does he knowingly permit his own services to be used by others, for purposes inconsistent with these values. While demanding for himself freedom of inquiry and communication, he accepts the responsibility this freedom confers: for competence where he claims it, for objectivity in the report of his findings, and for consideration of the best interests of his colleagues and of society.

SPECIFIC PRINCIPLES

Principle 1. Responsibility. The psychologist,[2] committed to increasing man's understanding of man, places high value on objectivity and integrity, and maintains the highest standards in the services he offers.

a. As a scientist, the psychologist believes that society will be best served when he investigates where his judgment indicates investigation is needed; he plans research in such a way as to minimize the possibility that his findings will be misleading; and he publishes full reports of his work, never discarding without explanation data which may modify the interpretation of results.
b. As a teacher, the psychologist recognizes his primary obligation to help others acquire knowledge and skill, and to maintain high standards of scholarship.
c. As a practitioner, the psychologist knows that he bears a heavy social responsibility because his work may touch intimately the lives of others.

---

[1] Copyrighted by the American Psychological Association, Inc., January 1963. Reprinted (and edited) from the *American Psychologist,* January 1963, and as amended by the APA Council of Representatives in September 1965 and December 1972. These standards are also reproduced in the *Biographical Directory of the American Psychological Association.*

[2] A student of psychology who assumes the role of psychologist shall be considered a psychologist for the purpose of this code of ethics.

Principle 2. Competence. The maintenance of high standards of professional competence is a responsibility shared by all psychologists, in the interest of the public and of the profession as a whole.

a. Psychologists discourage the practice of psychology by unqualified persons and assist the public in identifying psychologists competent to give dependable professional service. When a psychologist or a person identifying himself as a psychologist violates ethical standards, psychologists who know firsthand of such activities attempt to rectify the situation. When such a situation cannot be dealt with informally, it is called to the attention of the appropriate local, state, or national committee on professional ethics, standards, and practices.
b. Psychologists regarded as qualified for independent practice are those who (a) have been awarded a Diploma by the American Board of Professional Psychology, or (b) have been licensed or certified by state examining boards, or (c) have been certified by voluntary boards established by state psychological associations. Psychologists who do not yet meet the qualifications recognized for independent practice should gain experience under qualified supervision.
c. The psychologist recognizes the boundaries of his competence and the limitations of his techniques and does not offer services or use techniques that fail to meet professional standards established in particular fields. The psychologist who engages in practice assists his client in obtaining professional help for all important aspects of his problem that fall outside the boundaries of his own competence. This principle requires, for example, that provision be made for the diagnosis and treatment of relevant medical problems and for referral to or consultation with other specialists.
d. The psychologist in clinical work recognizes that his effectiveness depends in good part upon his ability to maintain sound interpersonal relations, that temporary or more enduring aberrations in his own personality may interfere with this ability or distort his appraisals of others. Therefore he refrains from undertaking any activity in which his personal problems are likely to result in inferior professional services or harm to a client; or, if he is already engaged in such an activity when he becomes aware of his personal problems, he seeks competent professional assistance to determine whether he should continue or terminate his services to his client.

Principle 3. Moral and Legal Standards. The psychologist in the practice of his profession shows sensible regard for the social codes and moral expectations of the community in which he works, recognizing that violations of accepted moral and legal standards on his part may involve his clients, students, or colleagues in damaging personal conflicts, and impugn his own name and the reputation of his profession.

Principle 4. Misrepresentation. The psychologist avoids misrepresentation of his own professional qualifications, affiliations, and purposes, and those of the institutions and organizations with which he is associated.

a. A psychologist does not claim either directly or by implication professional qualifications that differ from his actual qualifications, nor does he misrepresent his affiliation with any institution, organization, or individual, nor lead others to assume he has affiliations that he does not have. The psychologist is responsible for correcting others who misrepresent his professional qualifications or affiliations.

b. The psychologist does not misrepresent an institution or organization with which he is affiliated by ascribing to it characteristics that it does not have.
c. A psychologist does not use his affiliation with the American Psychological Association or its Divisions for purposes that are not consonant with the stated purposes of the Association.
d. A psychologist does not associate himself with or permit his name to be used in connection with any services or products in such a way as to misrepresent them, the degree of his responsibility for them, or the nature of his affiliation.

Principle 5. Public Statements. Modesty, scientific caution, and due regard for the limits of present knowledge characterize all statements of psychologists who supply information to the public, either directly or indirectly.

a. Psychologists who interpret the science of psychology or the services of psychologists to clients or to the general public have an obligation to report fairly and accurately. Exaggeration, sensationalism, superficiality, and other kinds of misrepresentation are avoided.
b. When information about psychological procedures and techniques is given, care is taken to indicate that they should be used only by persons adequately trained in their use.
c. A psychologist who engages in radio or television activities does not participate in commercial announcements recommending purchase or use of a product.

Principle 6. Confidentiality. Safeguarding information about an individual that has been obtained by the psychologist in the course of his teaching, practice, or investigation is a primary obligation of the psychologist. Such information is not communicated to others unless certain important conditions are met.

a. Information received in confidence is revealed only after most careful deliberation and when there is clear and imminent danger to an individual or to society, and then only to appropriate professional workers or public authorities.
b. Information obtained in clinical or consulting relationships, or evaluative data concerning children, students, employees, and others are discussed only for professional purposes and only with persons clearly concerned with the case. Written and oral reports should present only data germane to the purposes of the evaluation; every effort should be made to avoid undue invasion of privacy.
c. Clinical and other materials are used in classroom teaching and writing only when the identity of the persons involved is adequately disguised.
d. The confidentiality of professional communications about individuals is maintained. Only when the originator and other persons involved give their express permission is a confidential professional communication shown to the individual concerned. The psychologist is responsible for informing the client of the limits of the confidentiality.
e. Only after explicit permission has been granted is the identity of research subjects published. When data have been published without permission for identification, the psychologist assumes responsibility for adequately disguising their sources.
f. The psychologist makes provisions for the maintenance of confidentiality in the preservation and ultimate disposition of confidential records.

Principle 7. Client Welfare. The psychologist respects the integrity and protects the welfare of the person or group with whom he is working.

a. The psychologist in industry, education, and other situations in which conflicts of interest may arise among various parties, as between management and labor, or between the client and employer of the psychologist, defines for himself the nature and direction of his loyalties and responsibilities and keeps all parties concerned informed of these commitments.
b. When there is a conflict among professional workers, the psychologist is concerned primarily with the welfare of any client involved and only secondarily with the interest of his own professional group.
c. The psychologist attempts to terminate a clinical or consulting relationship when it is reasonably clear to the psychologist that the client is not benefiting from it.
d. The psychologist who asks that an individual reveal personal information in the course of interviewing, testing, or evaluation, or who allows such information to be divulged to him, does so only after making certain that the responsible person is fully aware of the purposes of the interview, testing, or evaluation and of the ways in which the information may be used.
e. In cases involving referral, the responsibility of the psychologist for the welfare of the client continues until this responsibility is assumed by the professional person to whom the client is referred or until the relationship with the psychologist making the referral has been terminated by mutual agreement. In situations where referral, consultation, or other changes in the conditions of the treatment are indicated and the client refuses referral, the psychologist carefully weighs the possible harm to the client, to himself, and to his profession that might ensue from continuing the relationship.
f. The psychologist who requires the taking of psychological tests for didactic, classification, or research purposes protects the examinees by insuring that the tests and test results are used in a professional manner.
g. When potentially disturbing subject matter is presented to students, it is discussed objectively, and efforts are made to handle constructively any difficulties that arise.
h. Care must be taken to insure an appropriate setting for clinical work to protect both client and psychologist from actual or imputed harm and the profession from censure.
i. In the use of accepted drugs for therapeutic purposes special care needs to be exercised by the psychologist to assure himself that the collaborating physician provides suitable safeguards for the client.

Principle 8. Client Relationship. The psychologist informs his prospective client of the important aspects of the potential relationship that might affect the client's decision to enter the relationship.

a. Aspects of the relationship likely to affect the client's decision include the recording of an interview, the use of interview material for training purposes, and observation of an interview by other persons.
b. When the client is not competent to evaluate the situation (as in the case of a child), the person responsible for the client is informed of the circumstances which may influence the relationship.
c. The psychologist does not normally enter into a professional relationship with members of his own family, intimate friends, close associates, or others whose welfare might be jeopardized by such a dual relationship.

Principle 9. Impersonal Services. Psychological services for the purpose of diagnosis, treatment, or personalized advice are provided only in the context of a professional relationship, and are not given by means of public lectures or demonstrations, newspaper or magazine articles, radio or television programs, mail, or similar media.

a. The preparation of personnel reports and recommendations based on test data secured solely by mail is unethical unless such appraisals are an integral part of a continuing client relationship with a company, as a result of which the consulting psychologist has intimate knowledge of the client's personnel situation and can be assured thereby that his written appraisals will be adequate to the purpose and will be properly interpreted by the client. These reports must not be embellished with such detailed analyses of the subject's personality traits as would be appropriate only after intensive interviews with the subject. The reports must not make specific recommendations as to employment or placement of the subject which go beyond the psychologist's knowledge of the job requirements of the company. The reports must not purport to eliminate the company's need to carry on such other regular employment or personnel practices as appraisal of the work history, checking of references, past performance in the company.

Principle 10. Announcement of Services. A psychologist adheres to professional rather than commercial standards in making known his availability for professional services.

a. A psychologist does not directly solicit clients for individual diagnosis or therapy.
b. Individual listings in telephone directories are limited to name, highest relevant degree, certification status, address, and telephone number. They may also include identification in a few words of the psychologist's major areas of practice, for example, child therapy, personnel selection, industrial psychology. Agency listings are equally modest.
c. Announcements of individual private practice are limited to a simple statement of the name, highest relevant degree, certification or diplomate status, address, telephone number, office hours, and a brief explanation of the types of services rendered. Announcements of agencies may list names of staff members with their qualifications. They conform in other particulars with the same standards as individual announcements, making certain that the true nature of the organization is apparent.
d. A psychologist or agency announcing nonclinical professional services may use brochures that are descriptive of services rendered but not evaluative. They may be sent to professional persons, schools, business firms, government agencies, and other similar organizations.
e. The use in a brochure of "testimonials from satisfied users" is unacceptable. The offer of a free trial of services is unacceptable if it operates to misrepresent in any way the nature or the efficacy of the services rendered by the psychologist. Claims that a psychologist has unique skills or unique devices not available to others in the profession are made only if the special efficacy of these unique skills or devices has been demonstrated by scientifically acceptable evidence.
f. The psychologist must not encourage (nor, within his power, even allow) a client to have exaggerated ideas as to the efficacy of services rendered. Claims made to clients about the efficacy of his services must not go beyond those which the psychologist would be willing to subject to professional scrutiny through publishing his results and his claims in a professional journal.

Principle 11. Interprofessional Relations. A psychologist acts with integrity in regard to colleagues in psychology and in other professions.

a. Each member of the Association cooperates with the duly constituted Committee on Scientific and Professional Ethics and Conduct in the performance of its duties by responding to inquiries with reasonable promptness and completeness. A member taking longer than 30 days to respond to such inquiries shall have the burden of demonstrating that he acted with "reasonable promptness."
b. A psychologist does not normally offer professional services to a person receiving psychological assistance from another professional worker except by agreement with the other worker or after the termination of the client's relationship with the other professional worker.
c. The welfare of clients and colleagues requires that psychologists in joint practice or corporate activities make an orderly and explicit arrangement regarding the conditions of their association and its possible termination. Psychologists who serve as employers of other psychologists have an obligation to make similar appropriate arrangements.

Principle 12. Remuneration. Financial arrangements in professional practice are in accord with professional standards that safeguard the best interest of the client and the profession.

a. In establishing rates for professional services, the psychologist considers carefully both the ability of the client to meet the financial burden and the charges made by other professional persons engaged in comparable work. He is willing to contribute a portion of his services to work for which he receives little or no financial return.
b. No commission or rebate or any other form of remuneration is given or received for referral of clients for professional services.
c. The psychologist in clinical or counseling practice does not use his relationships with clients to promote, for personal gain or the profit of an agency, commercial enterprises of any kind.
d. A psychologist does not accept a private fee or any other form of remuneration for professional work with a person who is entitled to his services through an institution or agency. The policies of a particular agency may make explicit provision for private work with its clients by members of its staff, and in such instances the client must be fully apprised of all policies affecting him.

Principle 13. Test Security. Psychological tests and other assessment devices, the value of which depends in part on the naivete of the subject, are not reproduced or described in popular publications in ways that might invalidate the techniques. Access to such devices is limited to persons with professional interests who will safeguard their use.

a. Sample items made up to resemble those of tests being discussed may be reproduced in popular articles and elsewhere, but scorable tests and actual test items are not reproduced except in professional publications.
b. The psychologist is responsible for the control of psychological tests and other devices and procedures used for instruction when their value might be damaged by revealing to the general public their specific contents or underlying principles.

Principle 14. Test Interpretation. Test scores, like test materials, are released only to persons who are qualified to interpret and use them properly.

a. Materials for reporting test scores to parents, or which are designed for self-appraisal purposes in schools, social agencies, or industry are closely supervised by qualified psychologists or counselors with provisions for referring and counseling individuals when needed.
b. Test results or other assessment data used for evaluation or classification are communicated to employers, relatives, or other appropriate persons in such a manner as to guard against misinterpretation or misuse. In the usual case, an interpretation of the test result rather than the score is communicated.
c. When test results are communicated directly to parents and students, they are accompanied by adequate interpretive aids or advice.

Principle 15. Test Publication. Psychological tests are offered for commercial publication only to publishers who present their tests in a professional way and distribute them only to qualified users.

a. A test manual, technical handbook, or other suitable report on the test is provided which describes the method of constructing and standardizing the test, and summarizes the validation research.
b. The populations for which the test has been developed and the purposes for which it is recommended are stated in the manual. Limitations upon the test's dependability, and aspects of its validity on which research is lacking or incomplete, are clearly stated. In particular, the manual contains a warning regarding interpretations likely to be made which have not yet been substantiated by research.
c. The catalog and manual indicate the training or professional qualifications required for sound interpretation of the test.
d. The test manual and supporting documents take into account the principles enunciated in the *Standards for Educational and Psychological Tests.*
e. Test advertisements are factual and descriptive rather than emotional and persuasive.

Principle 16. Research Precautions. The psychologist assumes obligations for the welfare of his research subjects, both animal and human.

The decision to undertake research should rest upon a considered judgment by the individual psychologist about how best to contribute to psychological science and to human welfare. The responsible psychologist weighs alternative directions in which personal energies and resources might be invested. Having made the decision to conduct research, psychologists must carry out their investigations with respect for the people who participate and with concern for their dignity and welfare. The Principles that follow make explicit the investigator's ethical responsibilities toward participants over the course of research, from the initial decision to pursue a study to the steps necessary to protect the confidentiality of research data. These Principles should be interpreted in terms of the contexts provided in the complete document[3] offered as a supplement to these Principles.

a. In planning a study the investigator has the personal responsibility to make a careful evaluation of its ethical acceptability, taking into account these Principles for research with human beings. To the extent that this appraisal, weighing scientific and humane values, suggests a deviation from any Principle, the investigator incurs an increasingly serious obligation to seek ethical advice and to observe more stringent safeguards to protect the rights of the human research participants.

---

[3] *Ethical Principles in the Conduct of Research with Human Participants,* available upon request from the American Psychological Association.

b. Responsibility for the establishment and maintenance of acceptable ethical practice in research always remains with the individual investigator. The investigator is also responsible for the ethical treatment of research participants by collaborators, assistants, students, and employees, all of whom, however, incur parallel obligations.

c. Ethical practice requires the investigator to inform the participant of all features of the research that reasonably might be expected to influence willingness to participate, and to explain all other aspects of the research about which the participant inquires. Failure to make full disclosure gives added emphasis to the investigator's abiding responsibility to protect the welfare and dignity of the research participant.

d. Openness and honesty are essential characteristics of the relationship between investigator and research participant. When the methodological requirements of a study necessitate concealment or deception, the investigator is required to ensure the participant's understanding of the reasons for this action and to restore the quality of the relationship with the investigator.

e. Ethical research practice requires the investigator to respect the individual's freedom to decline to participate in research or to discontinue participation at any time. The obligation to protect this freedom requires special vigilance when the investigator is in a position of power over the participant. The decision to limit this freedom gives added emphasis to the investigator's abiding responsibility to protect the participant's dignity and welfare.

f. Ethically acceptable research begins with the establishment of a clear and fair agreement between the investigator and the research participant that clarifies the responsibilities of each. The investigator has the obligation to honor all promises and commitments included in that agreement.

g. The ethical investigator protects participants from physical and mental discomfort, harm, and danger. If the risk of such consequences exists, the investigator is required to inform the participant of that fact, secure consent before proceeding, and take all possible measures to minimize distress. A research procedure may not be used if it is likely to cause serious and lasting harm to participants.

h. After the data are collected, ethical practice requires the investigator to provide the participant with a full clarification of the nature of the study and to remove any misconceptions that may have arisen. Where scientific or humane values justify delaying or withholding information, the investigator acquires a special responsibility to assure that there are no damaging consequences for the participant.

i. Where research procedures may result in undesirable consequences for the participant, the investigator has the responsibility to detect and remove or correct these consequences, including, where relevant, long-term aftereffects.

j. Information obtained about the research participants during the course of an investigation is confidential. When the possibility exists that others may obtain access to such information, ethical research practice requires that this possibility, together with the plans for protecting confidentiality, be explained to the participants as a part of the procedure for obtaining informed consent.

k. A psychologist using animals in research adheres to the provisions of the Rules Regarding Animals, drawn up by the Committee on Precautions and Standards in Animal Experimentation and adopted by the American Psychological Association.

l. Investigations of human subjects using experimental drugs (for example: hallucinogenic, psychotomimetic, psychedelic, or similar substances) should be conducted only in such settings as clinics, hospitals, or research facilities maintaining appropriate safeguards for the subjects.

Principle 17. Publication Credit. Credit is assigned to those who have contributed to a publication, in proportion to their contribution, and only to these.

a. Major contributions of a professional character, made by several persons to a common project, are recognized by joint authorship. The experimenter or author who has made the principal contribution to a publication is identified as the first listed.
b. Minor contributions of a professional character, extensive clerical or similar nonprofessional assistance, and other minor contributions are acknowledged in footnotes or in an introductory statement.
c. Acknowledgment through specific citations is made for unpublished as well as published material that has directly influenced the research or writing.
d. A psychologist who compiles and edits for publication the contributions of others publishes the symposium or report under the title of the committee or symposium, with his own name appearing as chairman or editor among those of the other contributors or committee members.

Principle 18. Responsibility toward Organization. A psychologist respects the rights and reputation of the institute or organization with which he is associated.

a. Materials prepared by a psychologist as a part of his regular work under specific direction of his organization are the property of that organization. Such materials are released for use or publication by a psychologist in accordance with policies of authorization, assignment of credit, and related matters which have been established by his organization.
b. Other material resulting incidentally from activity supported by any agency, and for which the psychologist rightly assumes individual responsibility, is published with disclaimer for any responsibility on the part of the supporting agency.

Principle 19. Promotional Activities. The psychologist associated with the development or promotion of psychological devices, books, or other products offered for commercial sale is responsible for ensuring that such devices, books, or products are presented in a professional and factual way.

a. Claims regarding performance, benefits, or results are supported by scientifically acceptable evidence.
b. The psychologist does not use professional journals for the commercial exploitation of psychological products, and the psychologist-editor guards against such misuse.
c. The psychologist with a financial interest in the sale or use of a psychological product is sensitive to possible conflict of interest in his promotion of such products and avoids compromise of his professional responsibilities and objectives.

APPENDIX B

Guidelines on Employee Selection Procedures (EEOC)[1]

Title 29—LABOR

Chapter XIV—Equal Employment Opportunity Commission

PART 1607—GUIDELINES ON EMPLOYEE SELECTION PROCEDURES

BY VIRTUE of the authority vested in it by section 713 of title VII of the Civil Rights Act of 1964, 42 U.S.C., section 2000e–12, 78 Stat. 265, the Equal Employment Opportunity Commission hereby issues Title 29, Chapter XIV, § 1607 of the Code of Federal Regulations.

These Guidelines on Employee Selection Procedures supersede and enlarge upon the Guidelines on Employment Testing Procedures, issued by the Equal Employment Opportunity Commission on August 24, 1966. Because the material herein is interpretive in nature, the provisions of the Administrative Procedure Act (5 U.S.C. 553) requiring notice of proposed rule making, opportunity for public participation, and delay in effective date are inapplicable. The Guidelines shall be applicable to charges and cases presently pending or hereafter filed with the Commission.

Sec.	
1607.1	Statement of purpose.
1607.2	"Test" defined.
1607.3	Discrimination defined.
1607.4	Evidence of validity.
1607.5	Minimum standards for validation.
1607.6	Presentation of validity evidence.
1607.7	Use of other validity studies.
1607.8	Assumption of validity.
1607.9	Continued use of tests.
1607.10	Employment agencies and employment services.
1607.11	Disparate treatment.
1607.12	Retesting.
1607.13	Other selection techniques.
1607.14	Affirmative action.

AUTHORITY: The provisions of this Part 1607 issued under Sec. 713, 78 Stat. 265, 42 U.S.C. sec. 2000e–12.

---

[1] From Federal Register, 1970, 35(149), 12333–12336. See also p. 63*n*, this book.

§ 1607.1 Statement of purpose.

(a) The guidelines in this part are based on the belief that properly validated and standardized employee selection procedures can significantly contribute to the implementation of nondiscriminatory personnel policies, as required by title VII. It is also recognized that professionally developed tests, when used in conjunction with other tools of personnel assessment and complemented by sound programs of job design, may significantly aid in the development and maintenance of an efficient work force and, indeed, aid in the utilization and conservation of human resources generally.

(b) An examination of charges of discrimination filed with the Commission and an evaluation of the results of the Commission's compliance activities has revealed a decided increase in total test usage and a marked increase in doubtful testing practices which, based on our experience, tend to have discriminatory effects. In many cases, persons have come to rely almost exclusively on tests as the basis for making the decision to hire, transfer, promote, grant membership, train, refer or retain, with the result that candidates are selected or rejected on the basis of a single test score. Where tests are so used, minority candidates frequently experience disproportionately high rates of rejection by failing to attain score levels that have been established as minimum standards for qualification.

It has also become clear that in many instances persons are using tests as the basis for employment decisions without evidence that they are valid predictors of employee job performance. Where evidence in support of presumed relationships between test performance and job behavior is lacking, the possibility of discrimination in the application of test results must be recognized. A test lacking demonstrated validity (i.e., having no known significant relationship to job behavior) and yielding lower scores for classes protected by title VII may result in the rejection of many who have necessary qualifications for successful work performance.

(c) The guidelines in this part are designed to serve as a workable set of standards for employers, unions and employment agencies in determining whether their selection procedures conform with the obligations contained in title VII of the Civil Rights Act of 1964. Section 703 of title VII places an affirmative obligation upon employers, labor unions, and employment agencies, as defined in section 701 of the Act, not to discriminate because of race, color, religion, sex, or national origin. Subsection (h) of section 703 allows such persons "* * * to give and to act upon the results of any professionally developed ability test provided that such test, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex or national origin."

§ 1607.2 "Test" defined.

For the purpose of the guidelines in this part, the term "test" is defined as any paper-and-pencil or performance measure used as a basis for any employment decision. The guidelines in this part apply, for example, to ability tests which are designed to measure eligibility for hire, transfer, promotion, membership, training, referral or retention. This definition includes, but is not restricted to, measures of general intelligence, mental ability and learning ability; specific intellectual abilities; mechanical, clerical and other aptitudes; dexterity and coordination; knowledge and proficiency; occupational and other interests; and attitudes, personality or temperament. The term "test" includes all formal, scored, quantified or standardized techniques of assessing job suitability including, in addition to the above, specific qualifying or disqualifying personal history or background requirements, specific educational or work history requirements, scored interviews, biographical information blanks, interviewers' rating scales, scored application forms, etc.

§ 1607.3 Discrimination defined.

The use of any test which adversely affects hiring, promotion, transfer or any other employment or membership opportunity of classes protected by title VII constitutes discrimination unless: (a) the test has been validated and evidences a high degree of utility as hereinafter described, and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer or promotion procedures are unavailable for his use.

§ 1607.4 Evidence of validity.

(a) Each person using tests to select from among candidates for a position or for membership shall have available for inspection evidence that the tests are being used in a manner which does not violate § 1607.3. Such evidence shall be examined for indications of possible discrimination, such as instances of higher rejection rates for minority candidates than nonminority candidates. Furthermore, where technically feasible, a test should be validated for each minority group with which it is used; that is, any differential rejection rates that may exist, based on a test, must be relevant to performance on the jobs in question.

(b) The term "technically feasible" as used in these guidelines means having or obtaining a sufficient number of minority individuals to achieve findings of statistical and practical significance, the opportunity to obtain unbiased job performance criteria, etc. It is the responsibility of the person claiming absence of technical feasibility to positively demonstrate evidence of this absence.

(c) Evidence of a test's validity should consist of empirical data demonstrating that the test is predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated.

(1) If job progression structures and seniority provisions are so established that new employees will probably, within a reasonable period of time and in a great majority of cases, progress to a higher level, it may be considered that candidates are being evaluated for jobs at that higher level. However, where job progression is not so nearly automatic, or the time span is such that higher level jobs or employees' potential may be expected to change in significant ways, it shall be considered that candidates are being evaluated for a job at or near the entry level. This point is made to underscore the principle that attainment of or performance at a higher level job is a relevant criterion in validating employment tests only when there is a high probability that

persons employed will in fact attain that higher level job within a reasonable period of time.

(2) Where a test is to be used in different units of a multiunit organization and no significant differences exist between units, jobs, and applicant populations, evidence obtained in one unit may suffice for the others. Similarly, where the validation process requires the collection of data throughout a multiunit organization, evidence of validity specific to each unit may not be required. There may also be instances where evidence of validity is appropriately obtained from more than one company in the same industry. Both in this instance and in the use of data collected throughout a multiunit organization, evidence of validity specific to each unit may not be required: *Provided,* That no significant differences exist between units, jobs, and applicant populations.

§ 1607.5 Minimum standards for validation.

(a) For the purpose of satisfying the requirements of this part, empirical evidence in support of a test's validity must be based on studies employing generally accepted procedures for determining criterion-related validity, such as those described in "Standards for Educational and Psychological Tests and Manuals" published by American Psychological Association, 1200 17th Street NW., Washington, D.C. 20036. Evidence of content or construct validity, as defined in that publication, may also be appropriate where criterion-related validity is not feasible. However, evidence for content or construct validity should be accompanied by sufficient information from job analyses to demonstrate the relevance of the content (in the case of job knowledge or proficiency tests) or the construct (in the case of trait measures). Evidence of content validity alone may be acceptable for well-developed tests that consist of suitable samples of the essential knowledge, skills or behaviors composing the job in question. The types of knowledge, skills or behaviors contemplated here do not include those which can be acquired in a brief orientation to the job.

(b) Although any appropriate validation strategy may be used to develop such empirical evidence, the following minimum standards, as applicable, must be met in the research approach and in the presentation of results which constitute evidence of validity:

(1) Where a validity study is conducted in which tests are administered to applicants, with criterion data collected later, the sample of subjects must be representative of the normal or typical candidate group for the job or jobs in question. This further assumes that the applicant sample is representative of the minority population available for the job or jobs in question in the local labor market. Where a validity study is conducted in which tests are administered to present employees, the sample must be representative of the minority groups currently included in the applicant population. If it is not technically feasible to include minority employees in validation studies conducted on the present work force, the conduct of a validation study without minority candidates does not relieve any person of his subsequent obligation for validation when inclusion of minority candidates becomes technically feasible.

(2) Tests must be administered and scored under controlled and standardized conditions, with proper safeguards to protect the security of test scores

and to insure that scores do not enter into any judgments of employee adequacy that are to be used as criterion measures. Copies of tests and test manuals, including instructions for administration, scoring, and interpretation of test results, that are privately developed and/or are not available through normal commercial channels must be included as a part of the validation evidence.

(3) The work behaviors or other criteria of employee adequacy which the test is intended to predict or identify must be fully described; and, additionally, in the case of rating techniques, the appraisal form(s) and instructions to the rater(s) must be included as a part of the validation evidence. Such criteria may include measures other than actual work proficiency, such as training time, supervisory ratings, regularity of attendance and tenure. Whatever criteria are used they must represent major or critical work behaviors as revealed by careful job analyses.

(4) In view of the possibility of bias inherent in subjective evaluations, supervisory rating techniques should be carefully developed, and the ratings should be closely examined for evidence of bias. In addition, minorities might obtain unfairly low performance criterion scores for reasons other than supervisors' prejudice, as, when, as new employees, they have had less opportunity to learn job skills. The general point is that all criteria need to be examined to insure freedom from factors which would unfairly depress the scores of minority groups.

(5) Differential validity. Data must be generated and results separately reported for minority and nonminority groups wherever technically feasible. Where a minority group is sufficiently large to constitute an identifiable factor in the local labor market, but validation data have not been developed and presented separately for that group, evidence of satisfactory validity based on other groups will be regarded as only provisional compliance with these guidelines pending separate validation of the test for the minority group in question. (See § 1607.9). A test which is differentially valid may be used in groups for which it is valid but not for those in which it is not valid. In this regard, where a test is valid for two groups but one group characteristically obtains higher test scores than the other without a corresponding difference in job performance, cutoff scores must be set so as to predict the same probability of job success in both groups.

(c) In assessing the utility of a test the following considerations will be applicable:

(1) The relationship between the test and at least one relevant criterion must be statistically significant. This ordinarily means that the relationship should be sufficiently high as to have a probability of no more than 1 to 20 to have occurred by chance. However, the use of a single test as the sole selection device will be scrutinized closely when that test is valid against only one component of job performance.

(2) In addition to statistical significance, the relationship between the test and criterion should have practical significance. The magnitude of the relationship needed for practical significance or usefulness is affected by several factors, including:

(i) The larger the proportion of applicants who are hired for or placed on the job, the higher the relationship needs to be in order to be practically useful. Conversely, a relatively low relationship may prove useful when proportionately few job vacancies are available;

(ii) The larger the proportion of applicants who become satisfactory employees when not selected on the basis of the test, the higher the relationship needs to be between the test and a criterion of job success for the test to be practically useful. Conversely, a relatively low relationship may prove useful when proportionately few applicants turn out to be satisfactory;

(iii) The smaller the economic and human risks involved in hiring an unqualified applicant relative to the risks entailed in rejecting a qualified applicant, the greater the relationship needs to be in order to be practically useful. Conversely, a relatively low relationship may prove useful when the former risks are relatively high.

§ 1607.6 Presentation of validity evidence.

The presentation of the results of a validation study must include graphical and statistical representations of the relationships between the test and the criteria, permitting judgments of the test's utility in making predictions of future work behavior. (See § 1607.5(c) concerning assessing utility of a test.) Average scores for all tests and criteria must be reported for all relevant subgroups, including minority and nonminority groups where differential validation is required. Whenever statistical adjustments are made in validity results for less than perfect reliability or for restriction of score range in the test or the criterion, or both, the supporting evidence from the validation study must be presented in detail. Furthermore, for each test that is to be established or continued as an operational employee selection instrument, as a result of the validation study, the minimum acceptable cutoff (passing) score on the test must be reported. It is expected that each operational cutoff score will be reasonable and consistent with normal expectations of proficiency within the work force or group on which the study was conducted.

§ 1607.7 Use of other validity studies.

In cases where the validity of a test cannot be determined pursuant to § 1607.4 and § 1607.5 (e.g., the number of subjects is less than that required for a technically adequate validation study, or an appropriate criterion measure cannot be developed), evidence from validity studies conducted in other organizations, such as that reported in test manuals and professional literature, may be considered acceptable when: (a) The studies pertain to jobs which are comparable (i.e., have basically the same task elements), and (b) there are no major differences in contextual variables or sample composition which are likely to significantly affect validity. Any person citing evidence from other validity studies as evidence of test validity for his own jobs must substantiate in detail job comparability and must demonstrate the absence of contextual or sample differences cited in paragraphs (a) and (b) of this section.

§ 1607.8 Assumption of validity.

(a) Under no circumstances will the general reputation of a test, its author or its publisher, or casual reports of test utility be accepted in lieu of evidence of validity. Specifically ruled out are: assumptions of validity based on test names or descriptive labels; all forms of promotional literature; data bearing on the frequency of a test's usage; testimonial statements of sellers, users, or consultants; and other nonempirical or anecdotal accounts of testing practices or testing outcomes.

(b) Although professional supervision of testing activities may help greatly to insure technically sound and nondiscriminatory test usage, such involvement alone shall not be regarded as constituting satisfactory evidence of test validity.

§ 1607.9 Continued use of tests.

Under certain conditions, a person may be permitted to continue the use of a test which is not at the moment fully supported by the required evidence of validity. If, for example, determination of criterion-related validity in a specific setting is practicable and required but not yet obtained, the use of the test may continue: *Provided:* (a) The person can cite substantial evidence of validity as described in § 1607.7 (a) and (b); and (b) he has in progress validation procedures which are designed to produce, within a reasonable time, the additional data required. It is expected also that the person may have to alter or suspend test cutoff scores so that score ranges broad enough to permit the identification of criterion-related validity will be obtained.

§1607.10 Employment agencies and employment services.

(a) An employment service, including private employment agencies, State employment agencies, and the U.S. Training and Employment Service, as defined in section 701(c), shall not make applicant or employee appraisals or referrals based on the results obtained from any psychological test or other selection standard not validated in accordance with these guidelines.

(b) An employment agency or service which is requested by an employer or union to devise a testing program is required to follow the standards for test validation as set forth in these guidelines. An employment service is not relieved of its obligation herein because the test user did not request such validation or has requested the use of some lesser standard than is provided in these guidelines.

(c) Where an employment agency or service is requested only to administer a testing program which has been elsewhere devised, the employment agency or service shall request evidence of validation, as described in the guidelines in this part, before it administers the testing program and/or makes referral pursuant to the test results. The employment agency must furnish on request such evidence of validation. An employment agency or service will be expected to refuse to administer a test where the employer or union does not supply satisfactory evidence of validation. Reliance by the test user on the reputation of the test, its author, or the name of the test shall not be deemed sufficient evidence of validity (see § 1607.8(a)). An employment agency or service may administer a testing program where the evidence of validity comports with the standards provided in § 1607.7.

§ 1607.11 Disparate treatment.

The principle of disparate or unequal treatment must be distinguished from the concepts of test validation. A test or other employee selection standard—

even though validated against job performance in accordance with the guidelines in this part—cannot be imposed upon any individual or class protected by title VII where other employees, applicants or members have not been subjected to that standard. Disparate treatment, for example, occurs where members of a minority or sex group have been denied the same employment, promotion, transfer or membership opportunities as have been made available to other employees or applicants. Those employees or applicants who have been denied equal treatment, because of prior discriminatory practices or policies, must at least be afforded the same opportunities as had existed for other employees or applicants during the period of discrimination. Thus, no new test or other employee selection standard can be imposed upon a class of individuals protected by title VII who, but for prior discrimination, would have been granted the opportunity to qualify under less stringent selection standards previously in force.

§ 1607.12 Retesting.

Employers, unions, and employment agencies should provide an opportunity for retesting and reconsideration to earlier "failure" candidates who have availed themselves of more training or experience. In particular, if any applicant or employee during the course of an interview or other employment procedure claims more education or experience, that individual should be retested.

§ 1607.13 Other selection techniques.

Selection techniques other than tests, as defined in § 1607.2, may be improperly used so as to have the effect of discriminating against minority groups. Such techniques include, but are not restricted to, unscored or casual interviews and unscored application forms. Where there are data suggesting employment discrimination, the person may be called upon to present evidence concerning the validity of his unscored procedures as well as of any tests which may be used, the evidence of validity being of the same types referred to in §§ 1607.4 and 1607.5. Data suggesting the possibility of discrimination exist, for example, when there are differential rates of applicant rejection from various minority and nonminority or sex groups for the same job or group of jobs or when there are disproportionate representations of minority and nonminority or sex groups among present employees in different types of jobs. If the person is unable or unwilling to perform such validation studies, he has the option of adjusting employment procedures so as to eliminate the conditions suggestive of employment discrimination.

§ 1607.14 Affirmative action.

Nothing in these guidelines shall be interpreted as diminishing a person's obligation under both title VII and Executive Order 11246 as amended by Executive Order 11375 to undertake affirmative action to ensure that applicants or employees are treated without regard to race, color, religion, sex, or national origin. Specifically, the use of tests which have been validated pursuant to these guidelines does not relieve employers, unions or employment agencies of their obligations to take positive action in affording employment and training to members of classes protected by title VII.

The guidelines in this part are effective upon publication in the FEDERAL REGISTER.

Signed at Washington, D.C., 21st day of July 1970.

[SEAL]

WILLIAM H. BROWN III,
Chairman.

[F.R. Doc. 70–9962; Filed, July 31, 1970; 8:46 A.M.]

Guidelines for Reporting Criterion-Related and Content Validity (OFCC)[1]

Chapter 60—Office of Federal Contract Compliance, Equal Employment Opportunity, Department of Labor

PART 60–3—EMPLOYEE TESTING AND OTHER SELECTION PROCEDURES

Guidelines for Reporting Criterion Related and Content Validity

ON February 14, 1973, notice of proposed rulemaking was published in the FEDERAL REGISTER (38 FR 4413) with regard to amending § 60–3.15 of Title 41, Code of Federal Regulations for the purpose of clarifying Federal contractors' existing obligations to report the validity of tests and other selection methods as required by 41 CFR Part 60–3. Interested persons were given thirty (30) days in which to submit written comments, suggestions, or objections regarding the proposed amendments.

Having considered all relevant materials submitted, I have determined that to effectuate the objectives of the proposal it would be more appropriate to amend 41 CFR 60–3.6, rather than 41 CFR 60–3.15 as originally proposed, since the purpose of the amendment is to clarify the manner in which evidence of validity is to be presented, and not to impose additional recordkeeping requirements.

For purposes of clarity, it has been determined that the different requirements for criterion-related validity and content validity should be stated separately. Specific requirements for construct validity studies are under consideration and may be added to 41 CFR § 60–3.6 at a later date.

---

[1] From *Federal Register,* 1974, *39*(12), 2094–2096.

Section 60–3.6 is hereby amended to read as follows:

§ 60–3.6 Presentation of evidence of validity.

(a) *Criterion-related validity.* Reports of criterion-related validity of tests and other selection methods, where required by this chapter, are to contain the following information:

(1) *Firm or organization and location(s) and date(s) of study.* If study is longitudinal, both testing dates and criterion collection dates or the time range within which these dates occurred must be shown (ESSENTIAL). If study was conducted at several locations, names of cities and states should be shown in addition to plant identification.

(2) *Problem and setting.* Explicit definition of purposes and brief discussion of the situation in which the study was conducted should be provided. Include description of existing selection procedures and cutting scores, if any, which were used.

(3) *Job analysis.* A full description of all important duties and tasks involved in each of the jobs on which research has been conducted and for which the test will be used must be provided (ESSENTIAL). Unless a general measure of job behavior (such as turnover or tenure) is used as a criterion, it is expected that the description of the duties performed by the incumbents on the job will also detail the level of difficulty at which the duties are performed, and those duties which are critical (i.e., must be performed competently if the job is to be performed in a satisfactory manner). This job analysis also should describe on what bases duties were determined to be critical, such as the proportion of time spent on the respective duties, their levels of difficulty, their importance, their frequency of performance, and other relevant factors. If a duty which occupies a high proportion of time is not deemed critical, the reasons therefore must be set forth. Published descriptions from industry sources or Volume I of the *Dictionary of Occupational Titles* are satisfactory if they adequately and completely described the job. If necessary a brief supplement to the published descriptions may be provided. (The requirements in this section are not intended to specify a particular method of job analysis. Any professionally recognized product and/or process oriented method of job analysis is acceptable if it is comprehensive and otherwise appropriate for the specific validation strategy used.) If two or more jobs are grouped for a validation study, a justification for this grouping as well as a description of each of the jobs should be provided.

(4) *Job titles and codes.* It is desirable that the job title used in company terminology plus the corresponding job title and code from United States Employment Service *Dictionary of Occupational Titles* (Third Edition) Volumes I and II, United States Government Printing Office, 1965, be provided. Where appropriate DOT codes and titles do not exist, a notation to that effect should be made.

(5) *Criteria.* Provide a description of all criteria of job related behaviors collected (including a rationale for what were selected as final criteria) and means by which they were observed, recorded, evaluated and quantified (ES-

SENTIAL). (To the extent possible this description should parallel the content validity justification described under "content validity" below.) An indication of the reliability coefficients and how they were established is desirable.

(6) *Sample.* A description of the ethnic and sex composition of the sample must be provided. (Minority group categories shown on EEO–1 should be used.) Also required is a description of how the research sample was selected and how it compares with employees in the job and with current applicants. Description of educational levels, length of service, and age is also desirable.

(7) *Number of cases.* Indicate the number of people in the sample, and the number of people in each of the subsamples (ESSENTIAL).

(8) *Predictors.* Commercially available tests should be described by title, form and publisher (ESSENTIAL). Other predictors, including privately constructed tests, interviewer's ratings, education and experience levels and composites of test batteries should be explicitly described (ESSENTIAL). The inclusion of a rationale for each predictor studied is desirable.

(9) *Techniques and results.* Methods used in evaluating data must be described (ESSENTIAL). Any statistical adjustments such as for less than perfect reliability or for restriction of score range in the test or criterion, or both must be described and uncorrected coefficients must also be shown (ESSENTIAL). Statistical results must be summarized in graphical or tabular form (by ethnic and/or subgroup, if feasible) and statements regarding statistical significance of results must be made (ESSENTIAL). Where the statistical technique used categorizes continuous data (such as biserial correlation and phi coefficient) the cut-offs used and their bases must be described (ESSENTIAL). Where more than one predictor and/or more than one criterion is used, all predictor-criterion correlations must be reported regardless of magnitude or direction. Average scores (e.g. means) and measures of dispersion (e.g. standard deviations and ranges) for all test and criteria must be reported for all relevant subgroups, including minority and nonminority groups and males and females, where feasible (ESSENTIAL).

(10) *Uses and applications.* A description of the way in which the test is to be used (e.g., as a screening device with a cut-off score, combined with other tests in a battery) and application of the test (e.g., selection, transfer, upgrading) must be provided so as to permit judgments of the test's utility in making predictions of future work behavior (ESSENTIAL). Where cut-off scores are to be used, both the cut-off scores and the way in which they were determined must be described (ESSENTIAL). (It is expected that each operational cut-off score will be reasonable and consistent with normal expectations of proficiency within the work force or group on which the study was conducted.) The specific way the test or test battery is used operationally must be described and validity data must support this particular use (ESSENTIAL). If predictor weights are used, the weights and the validity of the weighted composite must be reported (ESSENTIAL).

(11) *Source data.* Each contractor must maintain, and make available upon request, records showing all pertinent information about individual sample members in studies involving the validation of tests. These data must include test scores, criterion scores, age, sex, minority group status, and experience on the specific job on which the validation study is being conducted and may

also include such things as education, training and prior job experience. If the contractor chooses to submit along with a report on validation a worksheet showing the pertinent information about the individual sample members, specific identifying information such as name and social security number should NOT be shown. Submittal with the validation report is encouraged as it may prevent delays in approval in the event that errors are noted.

(12) *Contact person.* It would be desirable for the contractor to provide the name, address and telephone number of person currently working for the company or consulting firm who may be contacted for further information about the validation study.

(b) *Content validity.* Content validity is an appropriate type of validity to demonstrate for skills tests (that is, tests to determine whether an individual already possesses needed skills or knowledge), and certain other selection procedures, but it is not ordinarily an appropriate type of validity to fully support aptitude or intelligence tests (that is, tests to determine whether an individual can learn needed skills or knowledge). Also, content validity sometimes may be accepted as provisional compliance where criterion related is preferred but is not technically feasible. Frequently, the content validity of a selection procedure can be demonstrated to the satisfaction of the compliance officer by the contractor, either orally or in writing, without the need to prepare a formal report in accordance with these guidelines.

NOTE.—It is important to distinguish between face validity which involves an apparent but superficial correspondence between test content and the job (not an acceptable type of validity) and content validity, which involves a demonstration that the test content samples important elements of job tasks. Formal reports of the content validity of tests and other selection methods, where required by this chapter (i.e., disproportionate rejection rate and request of compliance officer), are to contain the following information:

(1) *Job analysis.* The essential foundation of a content validity study is a reliable and comprehensive job analysis. The analysis must define the duties performed by incumbents on the job, the level of difficulty at which the duties are performed, and those duties which are critical (i.e., those which must be performed competently if the job is to be performed in a satisfactory manner) (ESSENTIAL). The job analysis should describe on what bases duties were determined to be critical, such as the proportion of time spent on the respective duties, their levels of difficulty, their importance, their frequency of performance, and other relevant factors. If a duty which occupies a high proportion of time is not deemed critical, the reasons therefore must be set forth. (The level, breadth, degree and/or type of skills or knowledge to be learned on the job or in the job related training must be identified because selection standards must not require this level, breadth, degree and/or type of skills or knowledge.) Citation of actual occurrences of ineffective behavior resulting from the absence of a particular knowledge, skill or personal characteristic would be desirable.

(2) *Job title and code.* It is desirable that the job title used in company terminology plus the corresponding job title and code from United States Employment Service *Dictionary of Occupational Titles* (Third Edition) Volumes

I and II, United States Government Printing Office, 1965 be provided. Where appropriate DOT codes and titles do not exist, a notation to that effect should be made.

(3) *Predictors.* Privately constructed tests, specific training requirements, composites of test batteries and other predictors should be explicitly described (ESSENTIAL). Commercially available tests, if used, should be described by title, form and publisher (ESSENTIAL). (The predictor content must be restricted to measurement of the knowledge and skills identified in the job analysis.)

(4) *Techniques and results.* Describe the way in which the correspondence between the predictor content and the job content was established (ESSENTIAL). (This includes both content area and relative emphasis given to each area.) The validity demonstration must include a clear description of the depth and scope of primary characteristics required of the worker in performing these specific tasks. Specify the way in which these tasks, knowledge, skills and personal characteristics were determined and their relative importance. Describe qualifications of subject matter experts and/or job experts who reviewed predictor content and job requirements. If any steps were taken to reduce racial, cultural or sex bias in the predictor content, describe these steps. Explain how time limits were established and how these are related to speed with which tasks must be accomplished on the job. Describe studies, if any, to estimate the reliability of the predictors or, if relevant reliability studies have been described in the literature, these should be referenced.

(5) *Normative data.* As appropriate, it is desirable that average scores (e.g. means) and measures of dispersion (e.g. standard deviations and ranges) on the predictors be reported for one or more samples of applicants or other relevant samples. If feasible, these data should be reported separately for minority subgroups and separately for males and females.

(6) *Cut-off scores.* The rationale for minimum requirements, if any, must be provided (ESSENTIAL). (It is expected that each operational cut-off score or level of skill sought will be reasonable and consistent with normal expectations of proficiency within the work force.)

If the test is used to rank individuals who are above minimum levels, or if preference is given to individuals who score significantly above the minimum levels, a rationale for this procedure must be provided (ESSENTIAL).

(7) *Uses and applications.* A description of the way in which the test is to be used (e.g. as a screening device with a cut-off score, combined with other tests in a battery) and application of the test (e.g. selection, transfer, upgrading) must be provided (ESSENTIAL).

(8) *Contact person.* It would be desirable for the contractor to provide the name, mailing address and telephone number of person currently working for the company or consulting firm who may be contacted for further information about the validation study.

NOTE: Previously written company or consultant reports on criterion-related validity or content validity are acceptable if they are reasonably complete in regard to the above information. If the previously written reports are not complete, the additional information may be appended. If necessary information (all items except those identified as desirable) is not available, the report may still be submitted but approval will be limited to provisional compliance. Reports should not be submitted if any ESSENTIAL items are missing.

Effective date. This amendment shall become effective on January 17, 1974.

Signed at Washington, D.C., this 10th day of January 1974.

PETER J. BRENNAN,
Secretary of Labor.
BERNARD E. DELURY,
Assistant Secretary for Employment Standards.
PHILIP J. DAVIS,
Director, Office of Federal Contract Compliance.

FR Doc. 74-1461 Filed 1-16-74; 8:45 A.M.

APPENDIX C

A Suggested Outline for Test Evaluation

NOTE: The following outline contains basic features applicable to nearly all tests. Additional features are undoubtedly noteworthy in the case of particular tests. A good preparation for one's own evaluation of tests should include a thorough reading of the APA *Standards for Educational and Psychological Tests* and a perusal of a sample of test reviews in the *Mental Measurements Yearbooks.*

A. GENERAL INFORMATION

Title of test (including edition and forms if applicable)

Author(s)

Publisher, date of publication

Time required to administer

Cost (booklets, answer sheets, other test materials, available scoring services)

B. BRIEF DESCRIPTION OF PURPOSE AND NATURE OF TEST

General type of test (e.g., individual or group, performance, multiple aptitude battery, interest inventory)

Population for which designed (age range, type of person)

Nature of content (e.g., verbal, numerical, spatial, motor)

Subtests and separate scores

Type of items

C. PRACTICAL EVALUATION

Qualitative features of test materials (e.g., design of test booklet, editorial quality of content, ease of using, attractiveness, durability, appropriateness for examinees)

Ease of administration

Clarity of directions

Scoring procedures

Examiner qualifications and training

Face validity and examinee rapport

D. TECHNICAL EVALUATION

1. Norms

   Type (e.g., percentiles, standard scores)

   Standardization sample: nature, size, representativeness, procedures followed in obtaining sample, availability of subgroup norms (e.g., age, sex, education, occupation, region)

2. Reliability

   Types and procedure (e.g., retest, parallel-form, split-half, Kuder-Richardson), including size and nature of samples employed

   Scorer reliability if applicable

   Equivalence of forms

   Long-term stability when available

3. Validity

   Appropriate types of validation procedures (content, criterion-related predictive or concurrent, construct)

   Specific procedures followed in assessing validity and results obtained

   Size and nature of samples employed

E. REVIEWER COMMENTS

From *Mental Measurements Yearbooks* and other sources

F. SUMMARY EVALUATION

Major strengths and weaknesses of the test, cutting across all parts of the outline

APPENDIX D

Test Publishers

BELOW are the names and addresses of some of the larger American publishers and distributors of psychological tests. Catalogues of current tests can be obtained from these publishers on request. For names and addresses of other test publishers, see the Publisher's Directory in the latest *Mental Measurements Yearbook.*

Addison-Wesley Testing Service, 2725 Sand Hill Road, Menlo Park, Calif. 94025.
American Guidance Service, Inc., Publisher's Building, Circle Pines, Minn. 55014.
Bobbs-Merrill Company, Inc., 4300 West 62nd Street, Indianapolis, Ind. 46268.
Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa 52240.
CTB/McGraw-Hill, Del Monte Research Park, Monterey, Calif. 93940.
Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, Calif. 94306.
Educational and Industrial Testing Service, P.O. Box 7234, San Diego, Calif. 92107.
Educational Testing Service, Princeton, N.J. 08540.
Ginn and Company, 191 Spring Street, Lexington, Mass. 02173.
Houghton Mifflin Company, 110 Tremont Street, Boston, Mass. 02107.
Industrial Relations Center, University of Chicago, 1225 East 60th Street, Chicago, Ill. 60637.
Institute for Personality and Ability Testing, 1602 Coronado Drive, Champaign, Ill. 61820.
Psychological Corporation, a subsidiary of Harcourt Brace Jovanovich, Inc., 757 Third Avenue, New York, N.Y. 10017.
Psychological Test Specialists, Box 1441, Missoula, Mont. 59801.
Psychometric Affiliates, Box 3167, Munster, Ind. 46321.
Scholastic Testing Service, Inc., 480 Meyer Road, Bensenville, Ill. 60106.
Science Research Associates, Inc., 259 East Erie Street, Chicago, Ill. 60611.
Sheridan Psychological Services, P.O. Box 6101, Orange, Calif. 92667.
C. H. Stoelting Company, 1350 South Kostner Avenue, Chicago, Ill. 60623.
Western Psychological Services, 12031 Wilshire Boulevard, Los Angeles, Calif. 90025.

APPENDIX E

Classified List of Representative Tests

THIS table includes all published tests discussed in the text, except out-of-print tests cited for historical reasons only. In order to provide a more representative sample, a few additional tests not cited in the text are also listed under the appropriate categories. For more nearly complete coverage, the reader is referred to the *Mental Measurements Yearbooks* (MMY) and other sources cited in Chapter 1.

For each listed test, the table gives the volume and entry number in the MMY where the latest citation of that test can be found, including references to earlier reviews. When the latest edition of a test has not yet been listed in the MMY, reference to an earlier edition is given in parentheses, together with the publication date of the revised edition. This practice has also been followed in a few instances of substantial revisions in the manual, such as entirely new norms or extensive additions to validity data. For new tests, not yet included in the MMY, only the publication date is given.

For publishers not listed in Appendix D, the address is included in this table. It should be noted, however, that several of the major test publishers listed in Appendix D distribute tests in addition to those they publish. For convenience in ordering tests, the catalogues of those publishers should be consulted.

The tests are classified under the same categories employed in the text. All tests listed in this table are also included in the Subject Index.

Note: Effective January 1, 1976, several tests developed by Educational Testing Service are being published and distributed by Addison-Wesley Testing Service. Educational Testing Service will continue to be responsible for all professional and test-development aspects of these tests.

Note also that "Psychological Corporation, a subsidiary of Harcourt Brace Jovanovich, Inc.," is abbreviated throughout the table as "Psychological Corporation (HBJ)."

Title	Publisher	MMY (or date)
INTELLIGENCE TESTS AND DEVELOPMENTAL SCALES		
Individual Tests		
AAMD Adaptive Behavior Scale, 1974	American Association on Mental Deficiency, 5201 Connecticut Avenue, N.W., Washington, D.C., 20015	7–37
Arthur Adaptation of the Leiter International Performance Scale	Stoelting	6–526
Arthur Point Scale of Performance Tests:		4–335
Form 1	Stoelting	
Form 2	Psychological Corporation (HBJ)	
Bayley Scales of Infant Development	Psychological Corporation (HBJ)	7–402
Cain-Levine Social Competency Scale	Consulting Psychologists Press	6–69
Cattell Infant Intelligence Scale	Psychological Corporation (HBJ)	6–515
Columbia Mental Maturity Scale	Psychological Corporation (HBJ)	(6–517) 1972
Concept Assessment Kit—Conservation	Educational and Industrial Testing Service	7–437
Full Range Picture Vocabulary Test	Psychological Test Specialists	6–521
Gesell Developmental Schedules	Psychological Corporation (HBJ)	6–522
Haptic Intelligence Scale for Adult Blind	Psychology Research, Box 14, Technology Center, Chicago, Ill. 60616	7–409
Hiskey-Nebraska Test of Learning Aptitude	Marshall S. Hiskey, 5640 Baldwin, Lincoln, Neb. 68508	7–410
Leiter International Performance Scale	Stoelting	5–408
Lincoln-Oseretsky Motor Development Scale	Stoelting	5–767
McCarthy Scales of Children's Abilities	Psychological Corporation (HBJ)	1972
Oseretsky Tests of Motor Proficiency	American Guidance Service	4–650
Peabody Picture Vocabulary Test	American Guidance Service	7–417
Pictorial Test of Intelligence	Houghton Mifflin	7–418
Porteus Maze Test	Psychological Corporation (HBJ)	7–419
Quick Test	Psychological Test Specialists	7–422
Stanford-Binet Intelligence Scale	Houghton Mifflin	(7–425) 1973
Vineland Social Maturity Scale	American Guidance Service	6–194
Wechsler Adult Intelligence Scale (WAIS)	Psychological Corporation (HBJ)	7–429
Wechsler Intelligence Scale for Children—Revised (WISC-R)	Psychological Corporation (HBJ)	(7–431) 1974

Title	Publisher	MMY (or date)
INTELLIGENCE TESTS AND DEVELOPMENTAL SCALES (Cont.)		
Wechsler Preschool and Primary Scale of Intelligence (WPPSI)	Psychological Corporation(HBJ)	7–434
GROUP TESTS		
ACT Battery of the American College Testing Program	American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240	7–330
Alpha Examination, Modified Form 9	Psychological Corporation (HBJ)	3–220
Analysis of Learning Potential (ALP)	Psychological Corporation (HBJ)	7–334
California Test of Mental Maturity, 1963 Revision	CTB/McGraw-Hill	7–338
Cognitive Abilities Test	Houghton Mifflin	(7–343) 1971–1974
College Board Scholastic Aptitude Test (SAT)	Educational Testing Service (for the College Entrance Examination Board)	7–344
College Qualification Tests	Psychological Corporation (HBJ)	7–345
Concept Mastery Test	Psychological Corporation (HBJ)	(6–451) 1973
Culture Fair Intelligence Test	Institute of Personality and Ability Testing	6–453
D 48 Test	Consulting Psychologists Press	6–454
Goodenough-Harris Drawing Test	Psychological Corporation (HBJ)	7–352
Graduate Record Examinations (GRE)	Educational Testing Service (for the Graduate Record Board)	7–353; 7–667
Henmon-Nelson Tests of Mental Ability, 1973 Revision	Houghton Mifflin	(6–462) 1973
Kuhlmann-Anderson Tests: A Measure of Academic Potential	Ginn and Company	6–466
Miller Analogies Test	Psychological Corporation (HBJ)	7–363
Otis-Lennon Mental Ability Test	Psychological Corporation (HBJ)	7–370
Progressive Matrices	H. K. Lewis & Co. Ltd. (U.S. distributor: Psychological Corporation)	7–376
Revised Beta Examination	Psychological Corporation (HBJ)	6–494
School and College Ability Tests (SCAT)—Series II	Addison-Wesley Testing Service	(7–347) 1973
SRA Pictorial Reasoning Test	Science Research Associates	7–381
SRA Short Test of Educational Ability (STEA)	Science Research Associates	7–382
Tests of General Ability (TOGA)	Science Research Associates	6–496

Title	Publisher	MMY (or date)
MULTIPLE APTITUDE BATTERIES		
Academic Promise Tests (APT)	Psychological Corporation (HBJ)	7–672
ACT Career Planning Profile (CPP)	American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240	1971
Differential Aptitude Tests (DAT)	Psychological Corporation (HBJ)	(7–673) 1974
Flanagan Aptitude Classification Tests (FACT)	Science Research Associates	7–675
Flanagan Industrial Tests (FIT)	Science Research Associates	7–977
General Aptitude Test Battery (GATB)	U.S. Employment Service	7–676
Guilford-Zimmerman Aptitude Survey	Sheridan Psychological Services	6–772
Nonreading Aptitude Test Battery, 1969 Edition (NATB)	U.S. Employment Service	7–679
SRA Primary Mental Abilities, 1962 Edition	Science Research Associates	7–680
Structure-of-Intellect Abilities Tests (from Aptitudes Research Project, Guilford et al.)	Sheridan Psychological Services	*
TESTS OF CREATIVITY AND REASONING		
Creativity Tests for Children (from Aptitudes Research Project, Guilford et al.)	Sheridan Psychological Services	1974
Creativity-Verbal and Creativity-Nonverbal Tests (from Aptitudes Research Project, Guilford et al.)	Sheridan Psychological Services	*
Logical Reasoning	Sheridan Psychological Services	5–694
Pertinent Questions	Sheridan Psychological Services	6–557
Remote Associates Test	Houghton Mifflin	7–445
Ship Destination Test	Sheridan Psychological Services	6–500
Torrance Tests of Creative Thinking	Ginn and Company	7–448
Watson-Glaser Critical Thinking Appraisal	Psychological Corporation (HBJ)	7–883
EDUCATIONAL TESTS		
GENERAL ACHIEVEMENT BATTERIES		
Adult Basic Learning Examination (ABLE)	Psychological Corporation (HBJ)	7–3

---

* A few reviewed in MMY under individual titles.

Title	Publisher	MMY (or date)
EDUCATIONAL TESTS (Cont.)		
Basic Occupational Literacy Test (BOLT)	U.S. Employment Service	1973
California Achievement Tests, 1970 Edition	CTB/McGraw-Hill	7–5
Cooperative Primary Tests	Addison-Wesley Testing Service	7–10
Iowa Tests of Basic Skills	Houghton Mifflin	(6–13) 1974
Iowa Tests of Educational Development	Science Research Associates	(6–14) 1972
Fundamental Achievement Series	Psychological Corporation (HBJ)	7–978
McGraw-Hill Basic Skills System	McGraw-Hill Book Co., Inc., 1221 Avenue of the Americas, New York, N.Y. 10020	7–214, 230, 236, 477, 704, 781
Metropolitan Achievement Test	Psychological Corporation (HBJ)	7–14
Peabody Individual Achievement Test	American Guidance Service	7–17
Sequential Tests of Educational Progress (STEP)—Series II	Addison-Wesley Testing Service	(6–25) 1971
SRA Achievement Series	Science Research Associates	(7–18) 1971
SRA Reading and Arithmetic Indexes	Science Research Associates	7–20
Stanford Achievement Test	Psychological Corporation (HBJ)	(7–25) 1973
Stanford Test of Academic Skills (TASK)	Psychological Corporation (HBJ)	1973
Tests of Academic Progress (TAP)	Houghton Mifflin	(7–31) 1974
Wide Range Achievement Test, Revised Edition	Guidance Associates of Delaware, 1526 Gilpin Avenue, Wilmington, Delaware 19806	7–36
ACHIEVEMENT TESTS IN SEPARATE SUBJECTS		
Blyth Second-Year Algebra Test, Revised Edition	Psychological Corporation (HBJ)	7–497
Contemporary Mathematics Tests	CTB/McGraw-Hill	7–464
Content Evaluation Series:	Houghton Mifflin	
Language Arts Tests		7–198
Mathematics Test		7–475
Science Tests		7–799
Cooperative Mathematics Tests	Addison-Wesley Testing Service	7–465

Title	Publisher	MMY (or date)
EDUCATIONAL TESTS (Cont.)		
Cooperative Science Tests	Addison-Wesley Testing Service	7–787
MLA Cooperative Foreign Language Tests	Addison-Wesley Testing Service	7–254
Nelson Biology Test, Revised Edition	Psychological Corporation (HBJ)	7–819
Pimsleur Modern Foreign Language Proficiency Tests: French, German, Spanish	Psychological Corporation (HBJ)	7–257
Stanford Modern Mathematics Concepts Test	Psychological Corporation (HBJ)	7–490
DIAGNOSTIC AND CRITERION-REFERENCED TESTS		
Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic	Bobbs-Merrill	4–413
Diagnosis: An Instructional Aid Series—	Science Research Associates	
Mathematics		1972–73
Reading		1973–74
Diagnostic Reading Scales, Revised Edition	CTB/McGraw-Hill	(7–717) 1972
Diagnostic Reading Tests	Committee on Diagnostic Reading Tests, Inc., Mountain Home, N.C. 28758	6–823
Durrell Analysis of Reading Difficulty: New Edition	Psychological Corporation (HBJ)	5–660
Durrell Listening-Reading Series	Psychological Corporation (HBJ)	7–728
Gates-McKillop Reading Diagnostic Tests	Teachers College Press, 502 West 121st Street, New York, N.Y. 10027	6–824
Gilmore Oral Reading Test	Psychological Corporation (HBJ)	7–737
Gray Oral Reading Test	Bobbs-Merrill	6–842
Individual Pupil Monitoring System:	Houghton Mifflin	
Mathematics		1973
Reading		1974
Interim Evaluation Tests:	CTB/McGraw-Hill	
PMI		1973
PRI		1974
Iowa Silent Reading Tests	Psychological Corporation (HBJ)	(6–794) 1973
Nelson-Denny Reading Test	Houghton Mifflin	(6–800) 1973

Title	Publisher	MMY (or date)
EDUCATIONAL TESTS (Cont.)		
Prescriptive Mathematics Inventory	CTB/McGraw-Hill	1972
Prescriptive Reading Inventory	CTB/McGraw-Hill	1972
Skills Monitoring System:	Psychological Corporation (HBJ)	
Reading		1975–76
Study Skills		1976
Stanford Diagnostic Arithmetic Test	Psychological Corporation (HBJ)	7–528
Stanford Diagnostic Reading Test	Psychological Corporation (HBJ)	7–725
PROGNOSTIC TESTS		
Modern Language Aptitude Test	Psychological Corporation (HBJ)	6–357
Elementary Form		7–255
Orleans-Hanna Algebra Prognosis Test	Psychological Corporation (HBJ)	7–510
Orleans-Hanna Geometry Prognosis Test	Psychological Corporation (HBJ)	7–539
Pimsleur Language Aptitude Battery	Psychological Corporation (HBJ)	7–256
TESTS FOR EARLY CHILDHOOD EDUCATION		
Analysis of Readiness Skills: Reading and Mathematics	Houghton Mifflin	1972
Boehm Test of Basic Concepts	Psychological Corporation (HBJ)	7–335
Circus: An Assessment Program for Preprimary Children	Addison-Wesley Testing Service	1974
Cooperative Preschool Inventory—Revised Edition	Addison-Wesley Testing Service	7–404
Gesell Developmental Tests	Programs for Education, Box 85, Lumberville, Pa. 15933	7–750
Lee-Clark Reading Readiness Test	CTB/McGraw-Hill	7–752
Let's Look at Children	Addison-Wesley Testing Service	1975
Metropolitan Readiness Tests	Psychological Corporation (HBJ)	(7–757) 1976
Primary Reading Profiles, (Revised Edition)	Houghton Mifflin	5–665
Stanford Early School Achievement Test (SESAT)	Psychological Corporation (HBJ)	7–28
Tests of Basic Experiences (TOBE)	CTB/McGraw-Hill	7–33

Title	Publisher	MMY (or date)
OCCUPATIONAL TESTS		
SCREENING TESTS		
Otis Self-Administering Tests of Mental Ability	Psychological Corporation (HBJ)	5–363
Personnel Tests for Industry	Psychological Corporation (HBJ)	7–373
Wesman Personnel Classification Test	Psychological Corporation (HBJ)	7–400
Wonderlic Personnel Test	E. F. Wonderlic & Associates, P. O. Box 7, Northfield, Ill. 60093	7–401
SENSORY TESTS		
AO Sight Screener	American Optical Company, Southbridge, Mass. 01550	5–770
Dvorine Color Vision Test	Psychological Corporation (HBJ)	6–955
Keystone Telebinocular	Keystone View Company, Meadville, Pa. 16335	5–780
Massachusetts Hearing Test[1]	Massachusetts Department of Public Health, Division of Maternal and Child Health Services, 88 Broad Street, Boston, Mass. 02110	6–949
Ortho-Rater	Bausch & Lomb, Inc., Rochester, N.Y. 14602	5–783
School Vision Tester[1]	Bausch & Lomb, Inc., Rochester, N.Y. 14602	7–875
MOTOR TESTS		
Crawford Small Parts Dexterity Test	Psychological Corporation (HBJ)	5–871
Purdue Pegboard	Science Research Associates	6–1081
Stromberg Dexterity Test	Psychological Corporation (HBJ)	4–755
MECHANICAL APTITUDE TESTS		
Bennett Mechanical Comprehension Test	Psychological Corporation (HBJ)	7–1049
MacQuarrie Test for Mechanical Ability	CTB/McGraw-Hill	4–759
Revised Minnesota Paper Form Board Test	Psychological Corporation (HBJ)	7–1056
SRA Mechanical Aptitudes	Science Research Associates	4–764
CLERICAL TESTS		
Appraisal of Occupational Aptitudes	Houghton Mifflin	1971

---

[1] This test is designed for use with schoolchildren; it is listed here since this is the only section containing sensory tests.

Title	Publisher	MMY (or date)
OCCUPATIONAL TESTS (Cont.)		
General Clerical Test	Psychological Corporation (HBJ)	(4–730) 1972
Minnesota Clerical Test	Psychological Corporation (HBJ)	6–1040
Seashore-Bennett Stenographic Proficiency Test	Psychological Corporation (HBJ)	5–519
Short Employment Tests	Psychological Corporation (HBJ)	6–1045
Short Tests of Clerical Ability	Science Research Associates	6–1046
SRA Clerical Aptitudes	Science Research Associates	4–732
SRA Typing Skills (Forms A, B, C)	Science Research Associates	1975
Typing Tests for Business	Psychological Corporation (HBJ)	7–1007
ARTISTIC APTITUDE TESTS		
Graves Design Judgment Test	Psychological Corporation (HBJ)	4–220
Horn Art Aptitude Inventory	Stoelting	5–242
Meier Art Tests: 1. Art Judgment 2. Aesthetic Perception	Bureau of Educational Research and Service	7–240
MUSICAL APTITUDE TESTS		
Musical Aptitude Profile	Houghton Mifflin	7–249
Seashore Measures of Musical Talents	Psychological Corporation (HBJ)	6–353
Wing Standardized Tests of Musical Intelligence	NFER Publishing Co. Ltd., 2 Jennings Bldgs., Thames Avenue, Windsor, Berks. SL4 1QS, England	6–354
TESTS IN THE PROFESSIONS		
Dental Admission Testing Program	Division of Educational Measurements, Council on Dental Education, American Dental Association, 211 East Chicago Avenue, Chicago, Ill. 60611	7–1091
Law School Admission Test (LSAT)	Educational Testing Service	7–1098
Medical College Admission Test (MCAT)	Association of American Medical Colleges, One Dupont Circle, N.W., Washington, D.C. 20036	7–1100
Minnesota Engineering Analogies Test (MEAT)	Psychological Corporation (HBJ)	7–1095
National Teacher Examinations	Educational Testing Service	7–582
NLN Achievement Tests in Nursing	National League for Nursing, Inc., 10 Columbus Circle, New York, N.Y. 10019	7–1118, 7–1120

CLINICAL TESTS FOR COGNITIVE DYSFUNCTION AND LEARNING DISABILITIES

Title	Publisher	MMY (or date)
Auditory Discrimination Test (1973 Revision)	Western Psychological Services	(6–940) 1973
Bender-Gestalt Test	American Orthopsychiatric Association, Inc., 1790 Broadway, New York, N.Y. 10019	7–161
Concept Formation Test (Kasanin & Hanfmann)	Stoelting	6–78
Examining for Aphasia (Second Edition)	Psychological Corporation (HBJ)	5–52
Frostig Developmental Test of Visual Perception, Third Edition	Consulting Psychologists Press	7–871
Frostig Movement Skills Test Battery (Experimental Edition)	Consulting Psychologists Press	1973
Goldman-Fristoe-Woodcock Test of Auditory Discrimination	American Guidance Service	7–938
Goldstein-Scheerer Tests of Abstract and Concrete Thinking	Psychological Corporation (HBJ)	6–101
Illinois Test of Psycholinguistic Abilities, Revised Edition	University of Illinois Press, Urbana, Ill. 61801	7–442
Language Modalities Test for Aphasia	Industrial Relations Center, University of Chicago	7–956
Memory-for-Designs Test (Graham & Kendall)	Psychological Test Specialists	7–101
Minnesota Test for Differential Diagnosis of Aphasia	University of Minnesota Press, 2037 University Avenue, S.E., Minneapolis, Minn. 55455	7–958
Porch Index of Communicative Ability in Children	Consulting Psychologists Press	1974
Primary Visual Motor Test	Grune & Stratton, 757 Third Avenue, New York, N.Y. 10017	7–873
A Psychoeducational Inventory of Basic Learning Abilities	Fearon Publishers, 6 Davis Drive, Belmont, Calif. 94002	1968
The Pupil Rating Scale: Screening for Learning Disabilities	Grune & Stratton, 757 Third Avenue, New York, N.Y. 10017	1971
Revised Visual Retention Test (Benton)	Psychological Corporation (HBJ)	(6–543) 1974
Screening Tests for Identifying Children with Specific Language Disability, Revised Edition	Educators Publishing Service, Inc., 75 Moulton Street, Cambridge, Mass. 02138	7–969

Title	Publisher	MMY (or date)
CLINICAL TESTS FOR COGNITIVE DYSFUNCTION AND LEARNING DISABILITIES (Cont.)		
Southern California Sensory Integration Tests	Western Psychological Services	(6–63; 7–876 to 879) 1972
SELF-REPORT PERSONALITY INVENTORIES		
California Psychological Inventory (CPI)	Consulting Psychologists Press	7–49
Children's Personality Questionnaire	Institute for Personality and Ability Testing	(6–122) 1972
Clinical Analysis Questionnaire, Research Edition	Institute for Personality and Ability Testing	7–54 .
Comrey Personality Scales	Educational and Industrial Testing Service	7–59
Early School Personality Questionnaire	Institute for Personality and Ability Testing	7–71
Edwards Personal Preference Schedule (EPPS)	Psychological Corporation (HBJ)	7–72
Eysenck Personality Inventory	Educational and Industrial Testing Service	7–76
Gordon Personal Inventory	Psychological Corporation (HBJ)	6–102
Gordon Personal Profile	Psychological Corporation (HBJ)	6–103
Guilford-Zimmerman Temperament Survey	Sheridan Psychological Services	6–110
Jr.-Sr. High School Personality Questionnaire	Institute for Personality and Ability Testing	7–97
Minnesota Counseling Inventory	Psychological Corporation (HBJ)	6–142
Minnesota Multiphasic Personality Inventory (MMPI)	Psychological Corporation (HBJ)	7–104
Mooney Problem Check List	Psychological Corporation (HBJ)	6–145
Myers-Briggs Type Indicator	Consulting Psychologists Press	6–147
Personality Research Form	Research Psychologists Press, 13 Greenwich Avenue, Goshen, N.Y. 10924	7–123
Psychological Screening Inventory	Research Psychologists Press, 13 Greenwich Avenue, Goshen, N.Y. 10924	1973
Sixteen Personality Factor Questionnaire (16 PF)	Institute for Personality and Ability Testing	7–139
State-Trait Anxiety Inventory (STAI)	Consulting Psychologists Press	7–141
State-Trait Anxiety Inventory for Children (STAIC)	Consulting Psychologists Press	1973
STS Junior Inventory	Scholastic Testing Service	(5–104) 1968
STS Youth Inventory	Scholastic Testing Service	(6–170) 1971

Title	Publisher	MMY (or date)
SELF-REPORT PERSONALITY INVENTORIES (Cont.)		
Thorndike Dimensions of Temperament	Psychological Corporation (HBJ)	7–154
TESTS OF INTERESTS, ATTITUDES, AND VALUES		
California Occupational Preference Survey	Educational and Industrial Testing Service	7–1012
Career Maturity Inventory	CTB/McGraw-Hill	1973
Environmental Response Inventory (ERI)	Consulting Psychologists Press	1974
Gordon Occupational Check List	Psychological Corporation (HBJ)	7–1019
Guilford-Zimmerman Interest Inventory	Sheridan Psychological Services	6–1057
Hall Occupational Orientation Inventory	Scholastic Testing Service	(7–1021) 1971
Kuder General Interest Survey	Science Research Associates	7–1024
Kuder Occupational Interest Survey	Science Research Associates	7–1025
Kuder Personal Preference Record	Science Research Associates	6–132
Kuder Vocational Preference Record	Science Research Associates	6–1063
Minnesota Teacher Attitude Inventory	Psychological Corporation (HBJ)	6–699
Minnesota Vocational Interest Inventory (MVII)	Psychological Corporation (HBJ)	7–1026
Ohio Vocational Interest Survey (OVIS)	Psychological Corporation (HBJ)	7–1029
Personal Orientation Inventory	Educational and Industrial Testing Service	7–121
Pictorial Study of Values	Psychometric Affiliates	5–96
Self-Directed Search (SDS)	Consulting Psychologists Press	1971
Strong-Campbell Interest Inventory (SCII), 1974 edition of Strong Vocational Interest Blank (SVIB)	Stanford University Press, Stanford, Calif. 94305	(7–1036, 1037) 1974
Study of Values	Houghton Mifflin	7–146
Survey of Interpersonal Values	Science Research Associates	6–184
Survey of Personal Values	Science Research Associates	7–148
Survey of School Attitudes	Psychological Corporation (HBJ)	1975
Vocational Preference Inventory, Sixth Revision	Consulting Psychologists Press	7–157
Work Environment Preference Schedule	Psychological Corporation (HBJ)	1973
Work Values Inventory	Houghton Mifflin	7–1042
PROJECTIVE TECHNIQUES		
Animal Crackers: A Test of Motivation to Achieve	CTB/McGraw-Hill	1975

Title	Publisher	MMY (or date)
PROJECTIVE TECHNIQUES (Cont.)		
Blacky Pictures	Psychodynamic Instruments, P. O. Box 1221, Ann Arbor, Mich. 48106	6–204
Children's Apperception Test (CAT)	CPS Inc., P. O. Box 83, Larchmont, N.Y. 10538	(6–206) 1971
Gerontological Apperception Test	Behavioral Publications, Inc., 2852 Broadway, New York, N.Y. 10025	1971
Holtzman Inkblot Technique	Psychological Corporation (HBJ)	7–169
H-T-P: House-Tree-Person Projective Technique	Western Psychological Services	6–215
Kent-Rosanoff Free Association Test	Stoelting	6–226
Machover Draw-a-Person Test	Charles C Thomas, 327 East Lawrence Avenue, Springfield, Ill. 62703	6–229 (see also 7–165)
Make a Picture Story (MAPS)	Psychological Corporation (HBJ)	6–230
Piotrowski's Automated Rorschach (PAR)	Hay Associates, 1845 Walnut Street, Philadelphia, Pa. 19103	1974
Rorschach	Grune & Stratton, Inc., 757 Third Avenue, New York, N.Y. 10017 (U.S. distributor)	7–175
Rosenzweig Picture-Frustration Study (P-F Study)	Saul Rosenzweig, 8029 Washington Street, St. Louis, Mo. 63114	6–238
Rotter Incomplete Sentences Blank	Psychological Corporation (HBJ)	6–239
Senior Apperception Technique	CPS Inc., P. O. Box 83, Larchmont, N.Y. 10538	1973
Szondi Test	Grune & Stratton, Inc., 757 Third Avenue, New York, N.Y. 10017 (U.S. distributor)	6–243
Thematic Apperception Test (TAT)	Harvard University Press, 79 Garden Street, Cambridge, Mass. 02138	7–181
PERSONALITY TESTS: MISCELLANEOUS		
Adjective Check List (ACL)	Consulting Psychologists Press	7–38
Barron-Welsh Art Scale	Consulting Psychologists Press	7–41
Chapin Social Insight Test	Consulting Psychologists Press	7–51
Closure Flexibility (Concealed Figures)	Industrial Relations Center, University of Chicago	7–435
Closure Speed (Gestalt Completion)	Industrial Relations Center, University of Chicago	7–436
Embedded Figures Test	Consulting Psychologists Press	(6–89) 1973

Title	Publisher	MMY (or date)
PERSONALITY TESTS: MISCELLANEOUS (Cont.)		
Famous Sayings	Psychological Test Specialists	6–96
Fear Survey Schedule	Educational and Industrial Testing Service	7–80
IPAT Humor Test of Personality	Institute for Personality and Ability Testing	(4–61) 1966
MEASURES OF ENVIRONMENT		
American Home Scale	Psychometric Affiliates	5–596
College and University Environment Scales, Second Edition (CUES)	Educational Testing Service	7–56
The Inventory of College Activities	American Council on Education, One Dupont Circle, Washington, D.C. 20036	7–93
Social Climate Scales	Consulting Psychologists Press	1974
Stern Activities Index	National Computer Systems, 4401 West 76th Street, Minneapolis, Minn. 55435	7–142
Stern Environment Indexes (College Environment Index & others)	National Computer Systems, 4401 West 76th Street, Minneapolis, Minn. 55435	7–143

AUTHOR INDEX

References to bibliography have been set in bold face.

Abbott, E., 564, **648**
Abbott, M. M., 403, **679**
Abrahams, N. M., 180, **621**
Adams, H. L., 534, **657**
Adcock, C. J., 363, 585, **621**
Adkins, D. C., 415, 573, 574, **621**
Ahlström, K. G., 135, **621**
Airasian, P. W., 77 *n*, **624**
Albright, L. E., 149, **621**
Alexander, R. A., 194, **663**
Alf, E. F., Jr., 40, 180, **621, 642**
Ali, F., 285, **632**
Allen, J. M., 488, **645**
Allen, L., 329 *n*, 330, **648**
Allen, R. M., 284, **621**
Alley, W. E., 388, **675**
Allport, G. W., 508, 552, **621**
Amatruda, C. S., 76, 267, 268, **640**
Ames, L. B., 76, 428, 563, 579, **621, 622, 649**
Anastasi, A., 5 *n*, 46, 69, 100, 109 *n*, 110, 124, 125, 134, 226, 296, 326, 327, 338, 346, 351, 352, 353, 354, 377, 399 *n*, 436, 447, 522, 612, 615 *n*, **622, 623, 629, 634, 639, 655, 657, 673, 674, 678**
Anderson, J. E., 328, **623**
Anderson, R. P., 284, **666**
Anderson, S. B., 36, 267, 336, **623**
Angoff, W. H., 42, 43, 44, 91 *n*, 93, 94, 100, 222, 223, 318, 348, 411, **623**
Ansbacher, H. L., 295, **623**
Apell, R. J., 428, **649**
Appelbaum, M. I., 327, 329 *n*, 331, **659**
Apple, M. W., **661**
Arnold, G. F., 284, **623**
Arthur, A. Z., 488, **623**
Arthur, G., 289
Arvey, R. D., 610, **629**
Astin, A. W., 55, 617, 618, **623**
Atkinson, J. W., 356, 565, 567, 568, **624, 659, 675**
Atkinson, R. C., 96 *n*, **624**
Ayres, A. J., 482
Backer, T. E., 21, **632**
Bailey, D. E., 483, **668**
Bailey, J. P., Jr., 541, **634**
Bakan, D., 483, **624**
Bakare, C. G. M., 295, **624**
Bakeman, R., 615 *n*, **646**
Baker, C. T., 327, 329 *n*, 331, **650, 670**
Baker, F. B., 94, 95, 208, 302, 305, **624**
Ball, B., 504, **629**
Ball, R. S., 245, 272 *n*, 334, **671**
Baller, W. R., 339, **624**
Ballif, B. L., 573, 574, **621**
Balma, M. J., 150, 435, 609, **624, 655**
Baltes, P. B., 339, **624, 642, 678**
Bandura, A., 521, 523, **624**
Bannister, D., 603 *n*, **624**
Barker, R., 616, **624**
Barrett, R. S., 194, **653**
Barron, F., 389, 591, **624, 672**
Barron, R., 515, **649**
Bart, W. M., 77 *n*, **624**
Bartlett, C. J., 84, **625**
Bass, A. R., 194, **625**
Bass, B. M., 592, 593, 597, **625**
Bauernfeind, R. H., 540, **625**
Baughman, E. E., 504, 561, 562, 578, **625, 637**
Bayley, N., 272, 273, 329 *n*, 333, 334, 339, **625**
Bayroff, A. G., 388, **625**
Beattie, M., 557, **644**
Becker, W. C., 509, **625**
Bell, A., 339, **625**
Bell, F. O., 33, **625**
Bellak, L., 568, 569, **625**
Bellak, S. S., 569, **625**
Bemis, S. E., 386, **626**
Bender, L., 473, **626**
Bennett, G. K., 36, 451, **626**
Bentler, P. M., 77, 279, 280, 521, **626, 642**
Benton, A. L., 472, 473, **626**
Berdie, R. F., 179, 505 *n*, **626**
Berg, I. A., 520, 521, **625, 626, 648**

Bernardoni, L. C., 488, **666**
Bernberg, R., 559, **677**
Berner, J. G., 194, **668**
Bernstein, L., 34, **626**
Betz, N. E., 176, 302, 303, 304, **677**
Bianchini, J. C., 92, **657**
Biesheuvel, S., 297, **626**
Bijou, S. W., 604, **626**
Binet, A., 6, 7, 10, 11, 27, 74, 152, **626**
Bingham, W. V., 608 *n*, **626**
Birren, J. E., 341, **626**
Bishop, C. H., 44, **661**
Blake, R. H., 297, **626**
Blanchard, M. B., 222, **650**
Blin, Dr., 11 *n*
Block, J., 517 *n*, 520, 521, 601, 614, **626**, **627**
Bloom, B. S., 99, 136, 329, 352, 416, 618, **627**
Bloomberg, M., 389, **627**
Blumenkrantz, J., 339, **674**
Boersman, F. J., 44, **675**
Bolton, T. L., 10, **627**
Bommarito, J. W., 21, **650**
Bond, E. A., 244, **627**
Bonjean, C. M., 617, **627**
Booth, J. M., 615 *n*, **655**
Bordin, E. S., 542, **627**
Borgatta, E. F., 613, **656**
Boring, E. G., 5 *n*, **627**
Borislow, B., 519, **627**
Boruch, R. F., 55, **623**
Bouton, A., 537, **647**
Bower, E. M., 302, **627**
Bracht, G. H., **635**
Bradway, K. P., 327, 329 *n*, 330, **627**
Bratton, J. C., 189, **637**
Bray, D. W., 195, 196, 608 *n*, **627**, **643**
Breslow, E., 460, **627**
Bridgeman, B., 41, **628**
Briggs, P. F., 249, **672**
Brim, O. G., Jr., 51, 53, **667**
Brislin, R. W., 297, **628**
Brody, N., 335, **642**
Brogden, H. E., 172, 188, 190, **628**
Broman, S. H., 351, **663**
Brown, C. W., 172, 173, **628**
Brown, J. F., 573, **628**
Brown, S. R., 601, **628**
Brudner, H. J., 96, **638**
Buchholz, S., 284, **659**
Buchwald, A. M., 171, 172, **628**
Buech, B. U., 340, **668**
Burik, T. E., 139, **628**
Burke, H. R., 292, 293, **628**
Burnham, P. S., 164, 460, **628**
Burns, R. B., 339, **628**
Buros, O. K., 20, 502, **628**
Burt, C., 363, 375, 376, **628**
Burton, R. V., 595, **628**
Burtt, H. E., 570, **629**
Bush, W. J., 479 *n*, 480, **676**
Buss, A. R., 339, **629**
Butcher, J. N., 52, 502, 503, 504, **629**, **637**
Butler, B. V., 348, **641**

Caldwell, M. B., 348, **629**
Caldwell, O. W., 16, **629**
Campbell, D. P., 339, 342, 528, 530, 532, 533, 534, 535, **629**, **631**
Campbell, D. T., 156, 347, 559, 568, 576, 593, 607, **629**, **669**, **676**
Campbell, J. P., 610, **629**
Campbell, J. T., 194, 196, 436, **629**
Canfield, A. A., 83 *n*, **629**
Cannell, C. F., 608 *n*, **651**
Carkhuff, R. R., 541, **634**
Carnes, G. D., 541, **629**
Carroll, H. A., 454, **629**
Carroll, J. B., 99, 377, **629**, **630**
Carter, L. F., 597, **630**
Cattell, J. McK., 9, 569, **630**
Cattell, R. B., 289, 508, 509, 510, 589, 592, **630**, **657**
Cautela, J. R., 606, **630**
Chaille, S. E., 11 *n*
Chapman, J. P., 485, 582, **630**
Chapman, L. J., 485, 581, 582, **630**
Charles, D. C., 339, 342, **624**, **630**
Chase, C. I., **655**
Chave, E. J., 546, 547, 548, 550, **674**
Christal, R. E., 372, **630**
Chun, Ki-Taek, 21, **630**
Clark, J. H., 504, **631**
Clark, K. E., 535, **631**
Clark, R. A., 568, **659**
Cleary, T. A., 57, 176, 192, 195, 196, 223, 303, **631**, **656**
Clemans, W. V., 32, 538 *n*, **631**
Cobb, S., 21, **630**
Coffman, W. E., **645**
Cohen, E., 39, **631**
Cohen, J., 254, **631**
Cohen, L. D., 250, **636**
Cohen, R. A., 347, **631**

Cole, N. S., 197, 631
Cole, S., 348, 631
Coleman, W., 401, 631
Colgin, R., 504, 640
Collins, B. E., 557, 631
Collins, M. G., 284, 621
Combs, A. W., 598, 670
Comrey, A. L., 21, 632
Conrad, H. S., 222, 650
Cooley, W. W., 92, 96, 99, 632
Corah, M. L., 519, 632
Cordova, F. A., 346, 622
Corrigan, B., 485, 486, 633
Costello, J., 285, 632
Couch, A., 520, 521, 632
Courtis, S. A., 16, 629
Cox, J. A., Jr., 613, 637
Cozby, P. C., 525, 632
Craik, K. H., 616, 617, 632
Cravens, R. B., 327, 627
Crawford, A. B., 460, 628
Creager, J. A., 322, 632
Crites, J. O., 435, 672
Cronbach, L. J., 118, 124 *n*, 161, 169, 177 *n*, 205, 297, 483, 580, 586, 624, 626, 632, 640, 663, 674
Crooks, L. A., 194, 436, 622, 629
Crowne, D. P., 516, 632
Cruikshank, R. M., 451, 626
Crumrine, W. M., 537, 658
Cundick, B. P., 285, 632
Cureton, E. E., 115 *n*, 172, 208, 220, 221, 401, 631, 632, 633

Dahlstrom, L. E., 499, 500, 501, 502, 503, 633
Dahlstrom, W. G., 499, 500, 501, 502, 503, 504, 625, 633
Dailey, C. A., 483, 633
Dailey, J. T., 91
Damaye, H., 11 *n*
Damon, A., 339, 633
Damrin, D. E., 16, 635
D'Andrade, R. G., 509, 633
Daniels, C. W., 504, 661
Darley, J. G., 536, 541, 542, 633
Darlington, R. B., 169, 633
Das, R. S., 346, 633
Davids, A., 577, 633
Davidson, K. S., 37, 38, 667, 676
Davis, A., 329, 627
Davis, P. C., 254, 633
Davis, W. E., 40, 633
Dawes, R. M., 485, 486, 633
Day, G. J., 541, 668
Dearborn, W. F., 43, 633
Delys, P., 557, 671
Demming, J. A., 342, 633
Denenberg, V. H., 377, 677
Dennis, W., 295, 633
Denny, J. P., 38, 633
Deutsch, M., 58, 61, 633, 650
DeWitt, L. J., 176, 303, 304 *n*, 633
Diaz-Guerrero, R., 295, 655
Dicken, C. F., 519, 634
Dickson, C. R., 604, 634
Doll, E. A., 268, 269, 634
Donofrio, A. F., 267, 333, 634
Doppelt, J. E., 36, 249, 250, 255, 259, 263, 337, 626, 634, 669
Dorcus, R. M., 440, 451, 634
Drake, J., 109 *n*, 110, 124, 125, 622
Drake, L. E., 501, 634
Drasgow, J., 541, 634
Dreger, R. M., 354, 634
Drenth, P. J. D., 297, 624, 626, 632, 663, 674
Drews, E. M., 348, 672
Droege, R. C., 44, 386, 634
DuBois, P. H., 5, 18, 187, 318, 634
Duke, M. P., 557, 663
Duker, J., 501, 640
Dunn, J. A., 294, 634
Dunnette, M. D., 149, 180, 489 *n*, 610, 629, 634, 635, 647
Dunteman, G. H., 541, 635
Durnall, E. J., Jr., 537, 635
Dvorak, B. J., 384, 635
Dyer, H. S., 39, 635
Dyk, R. B., 591, 678
Dymond, R. F., 601, 666
D'Zurilla, T. J., 523, 641

Ebbinghaus, H., 10, 635
Ebel, R. L., 16, 44, 97, 100, 117 *n*, 118, 135 *n*, 149, 204, 211, 212, 399 *n*, 415, 635, 661
Eber, H. W., 509, 630
Edgerton, H. A., 84, 214, 625, 635
Edwards, A. J., 243, 635
Edwards, A. L., 516, 517, 519, 551, 635
Eells, K., 225, 345, 617, 636, 676
Eisdorfer, C., 250, 339, 636
Ekstrom, R. B., 373, 639
Elliott, L. L., 478 *n*, 672
Ellis, N. R., 666

Ellison, C. D., 445, **638**
Ellison, R. L., 615 *n*, **650**
Endler, N. S., 179, 523, **636**, **643**
Engelhart, M. D., 211, **636**
Entwisle, D. R., 567, **636**
Epstein, S., 584, **636**
Eriksen, C. W., 38, **664**
Erikson, E. H., 334, **636**
Ernhart, C. B., 468, 469, **643**
Eron, L. D., 558, **679**
Escalona, S. K., 277 *n*, 333, **636**
Esquirol, J. E. D., 6, **636**
Estes, B. W., 259, **636**
Evans, F. R., 43, **664**
Evans, I. M., 43, 604, **636**
Ewen, R. B., 194, **653**
Exner, J. E., Jr., 39, **636**

Fan, C. T., 201, 214, **636**, **637**
Farley, J., 583, 587, **646**
Farnsworth, P. R., 550, **637**
Fateson, H. F., 591, **678**
Fear, R. A., 608 *n*, **637**
Feather, N. T., 568, **624**
Featherman, D. L., **624**
Feder, D. D., 599, **671**
Feld, S. C., 567, **675**
Feldhusen, J. F., 38, **637**
Feldman, D. H., 189, **637**
Felling, J., 575, **676**
Ferguson, G. A., 377, **637**
Ferguson, R. L., 132, 133, **637**
Ferrari, G. C., 10, **643**
Field, J. G., 466, **637**
Figurelli, J. C., 280, **637**
Fincher, C., 62 *n*, **637**
Findley, W. G., 212, **637**
Fink, A. M., 52, **637**
Fishbein, M., 545 *n*, 551, **637**
Fisher, J., 164, **637**
Fishman, J. A., 58, 550, 633, **637**
Fiske, D. W., 156, 157, 515, 562, 597, 613, 629, **637**, **652**
Fitzgibbon, T. J., 59, **637**
Fjeld, S. P., 604, **638**
Flanagan, J. C., 91, 96, 100, 102, 437, 438, 598, **638**
Flamer, G. B., 77, **643**
Flaugher, R. L., 58, **638**
Flavell, J. H., 77, **638**
Fleishman, E. A., 148, 377, 439, 443, 445, 446, **638**, **673**
Force, R. C., 541, **638**
Ford, M. P., 77, **643**
Ford, S. F., 222, 223, **623**
Forer, B., 489 *n*, **639**
Fossum, J. A., 216, **639**
Fox, D. G., 615 *n*, **650**
Franks, C. M., 523, **639**, **651**
Frederiksen, C. H., 377, **639**
Frederiksen, N., 179, 438, 516, 616, **639**
Freeberg, N. E., 610, **639**
Freeman, M., 329 *n*, 331, **650**
Freitag, G., 608, **657**
French, J. R. P., Jr., 21, **630**
French, J. W., 38, 190, 372, 373, 377, **639**
Frengel, B. A., 396, **679**
Friedlander, B. Z., **649**
Fromm, E., 559, **639**
Fromme, A., 573, **639**
Frostig, M., 482 *n*
Fruchter, B., 83 *n*, 104, 148, 182, 208, 215 *n*, **638**, **644**
Fryer, D., 528, **639**
Fuchs, E. F., 189, 194, 196, 388, **625**, **658**
Fulkerson, S. C., 179, **640**

Gage, N. L., 483, **640**
Gagné, R., 99, **640**
Galton, F., 8, 9, 569, **640**
Gamble, K. R., 564, **640**
Gardner, J. W., 61, **640**
Gaudry, E., 37, **640**
Georgopoulos, B. S., 149, **669**
Gerberich, J. R., 410, 617 *n*, **640**
Gesell, A., 76, 267, 268, 334, **640**
Ghiselli, E. E., 147, 148, 164, 172, 173, 179, 180, 446, 448, 628, **640**, **651**
Gilberstadt, H., 501, **640**
Gilbert, A. C. F., 179, **639**
Gilbert, H. B., 61, **640**
Gilbert, J. A., 10, **640**
Gilliland, A. R., 504, **640**
Ginsburg, H., 77, **640**
Gladwin, T., 469, **667**
Glaser, E. M., 21, **632**
Glaser, R., 96, 99, 132, 133, 302, 305, 421, **632**, **638**, **641**
Glasser, A. J., 246, 466, 467, 470, 488, **641**, **679**
Glennon, J. R., 149, **621**
Gleser, G. C., 169, 177 *n*, 586, **632**
Goffeney, B., 348, **641**
Gold, V. J., 608, **657**

Goldberg, L. R., 18, 485, 486, 515, 641, 645
Golden, M., 583, 641
Goldfried, M. R., 521, 523, 524 *n*, 562, 604, 641
Golding, S. L., 485, 582, 641
Goldman, L., 57, 435, 641
Goldschmid, M. L., 77, 279, 280, 281, 641, 642
Goldsmith, J. A., 524, 642
Goldstein, K., 467, 476
Goodenough, D. R., 591, 678
Goodenough, F. L., 5 *n*, 11 *n*, 295, 642
Goodstein, L. D., 493, 655
Gordon, E., 458, 642
Gordon, E. W., 329, 642
Gordon, L. V., 40, 552, 642
Gordon, M. A., 193, 196, 642
Gorham, D. R., 564, 642
Goslin, D., 653
Gottfried, A. W., 335, 642
Gough, H. G., 225, 600, 613, 617, 642
Goulet, L. R., 339, 642, 678
Graham, F. K., 468, 469, 642, 643
Grant, D. L., 196, 608 *n*, 643
Gravem, H., 334, 649
Gray, F. E., 518, 669
Gray, S. W., 285, 653
Green, B. F., Jr., 305, 643
Green, D. R., 77, 399 *n*, 643
Green, R. F., 516, 643
Greene, H. A., 410, 617 *n*, 640
Greene, M. M., 542, 674
Greenwood, J. M., 597, 643
Greif, E. B., 555, 654
Gronlund, N. E., 97, 403, 413, 414, 415, 643
Grooms, R. R., 179, 643
Gross, A. L., 197, 643
Grossman, H., 238, 239, 643
Gruenfeld, L. W., 591, 677
Guertin, W. H., 246, 249, 253, 254, 465, 466, 468, 643, 665
Guicciardi, G., 10, 643
Guilford, J. P., 83 *n*, 104, 182, 208, 215 *n*, 363, 366, 374, 375, 376, 389, 391, 393, 394, 396, 397, 449, 506, 541, 562, 596, 597, 644
Guion, R. M., 150, 151, 644
Gulliksen, H., 120 *n*, 124 *n*, 129, 196, 551, 644
Gurin, G., 557, 567, 644, 675
Gurin, P., 557, 644
Gustad, J., 608, 626
Guttman, I., 169, 644
Guttman, L., 76, 124, 644
Gynther, M. D., 504, 659

Haan, N., 329 *n*, 331, 644
Haddad, J., 522, 545, 593, 654
Hagen, E., 311, 312, 313
Hagenah, T., 536, 541, 542, 633
Haire, M., 148, 640
Hall, J., 253, 644
Haller, D. L., 501, 658
Halverson, H. M., 76, 644
Hambleton, R. K., 96, 132, 133, 644, 645
Hamilton, R. G., 578, 645
Hammond, K. R., 485, 488, 645
Hansen, J.-I. C., 541, 645
Harari, O., 610, 645
Harlow, H. F., 377, 645
Harman, H. H., 363, 645
Härnqvist, K., 327, 330, 331, 342, 645
Harris, C. W., 135, 618, 645, 653
Harris, D. B., 293, 294, 295, 642, 645
Harris, J. A., 95, 96, 645
Harrison, R. H., 504, 565, 568, 583, 645
Hartman, L. D., 559, 639
Hartshorne, H., 19, 594, 595, 613, 645
Hase, H. D., 515, 645
Hata, Y., 34, 674
Hathaway, S. R., 497, 501, 646
Hauser, R. M., 624
Havighurst, R. J., 253, 334, 646, 650
Hay, E. N., 451, 646
Hayes, K. J., 355, 356, 646
Hayes, S. P., 283, 646
Hebb, D. O., 351, 354, 646
Heilbrun, A. B., Jr., 521, 600, 614, 642, 646
Heim, A. W., 44, 646
Hellervik, L. V., 610, 629
Helmreich, R., 615 *n*, 646
Helmstadter, G. C., 124 *n*, 646
Hempel, W. E., Jr., 148, 445, 638
Henderson, N. B., 348, 641
Henri, V., 10, 626
Henry, E. R., 615 *n*, 663
Henry, W. E., 565, 567, 583, 587, 646
Henrysson, S., 208, 216, 646
Herman, P. S., 580, 656
Herron, E. W., 563, 564, 578, 646, 647, 648
Herskovits, M. J., 347, 669

Hertzman, M., **678**
Herzberg, F., 537, **647**
Hess, R., 329, **627**
Hester, R., 537, **662**
Hewer, V. H., 196, **647**
Hicks, L. E., 512, **647**
Hickson, R., 606, **672**
Higgins, C., 348, **647**
Hilkert, R. N., 33, **674**
Hill, E. F., 564, 565, **647**
Hill, K. T., 37, **647**, **667**
Hill, R. J., 617, **627**
Hilton, T. L., 223, **631**
Himelstein, P., 234, 244, **647**
Hinckley, E. D., 550, **647**
Hobert, R., 180, **647**
Hoepfner, R., 366, 374, 375, 393, **644**
Hoff, A. L., 33, **625**
Hoffman, B., 302, **647**
Hoffman, P. J., 169, 485, **666**, **678**
Hogan, T. P., 403, **679**
Hogarty, P. S., 327, 329 *n*, 331, 333, **659**
Holden, R. H., 284, **647**
Holland, J. L., 530, 535, 536, 542, 552, **647**
Hollander, E. P., 613, **647**
Hollenbeck, G. P., 264, 275, **647**, **651**
Hollingshead, A. B., 617, **647**
Holmen, M. G., 597, **647**
Holt, R. R., 488, **647**
Holtzman, W. H., 52, 95, 552, 563, 564, **643**, **647**, **648**, **657**
Holzberg, J. D., 558, **648**
Honzik, M. P., 329 *n*, 330, 334, **648**, **677**
Hooper, F. H., 279 *n*, **648**
Hopkins, K. D., 415, 617 *n*, **635**, **671**
Hopkins, T., 410, **675**
Horn, C. C., 455, **648**
Horst, P., 190, **648**
House, J. S., 617, **655**
Hovland, C. I., 550, **648**
Howard, J. L., 285, **648**
Hoyt, K. B., 33, **625**
Hsieh, Kuo-Cheng, 169, **666**
Huber, J. T., 488, **648**
Huddleston, E. M., 135 *n*, **648**
Hughes, R. B., 348, **648**
Hull, C. L., 437, **648**
Humphreys, L. G., 57, 192, 194, 375, 399 *n*, **631**, **648**, **649**
Hunt, J., McV., 77, 278, 279, 425, 431, 523, **636**, **649**, **663**, **675**
Hunter, J. E., 194, 197, **668**
Hunter, M., 348, **631**
Hurlburt, N., 333, **659**
Hursch, C. J., 485, **645**
Hurvich, M. S., 569, **625**
Husek, T. R., 98 *n*, 132 *n*, **665**
Husén, T., 327, 342, **649**

Ilg, F. L., 428, **649**
Imparato, N., 610, **679**
Indik, B. P., 149, **669**
Inhelder, B., 277 *n*
Ireton, H., 334, **649**
Irvine, S. H., 346, **649**

Jackson, D. N., 513, 514, 515, 517 *n*, 520, 521, 541, **626**, **649**, **669**
Jacobs, A., 515, **649**
Jacobsen, T. L., 149, **666**
Jaeger, R. M., 92, **649**
James, L. R., 160, 615 *n*, **650**
James, S. T., 339, 342, **630**
Janke, L. L., 253, **650**
Jarrett, R. F., 172, **650**
Jastak, J. F., 480
Jastrow, J., 9
Jeanneret, P. R., 439, **659**
Jenkins, J. J., 571, **650**, **663**
Jensen, A. R., 346, 350, 351, **650**
Johansson, C. B., 541, **645**
Johnson, A. P., 211, **650**
Johnson, D. J., 479 *n*, **650**
Johnson, O. G., 21, **650**
Johnson, R. W., 488, **666**
Jones, H. E., 222, **650**
Jones, H. G., 466, **650**
Jones, L. V., 245, **650**
Jones, M. C., 523, **650**
Jones, M. H., 440, 451, **634**
Jones, R. R., 607, 608, **650**
Jorgensen, A. N., 410, 617 *n*, **640**
Jung, C. G., 569, 570, **650**
Jurgensen, C. E., 213, 518, **650**

Kagan, J., 329 *n*, 331, **650**
Kahn, R. L., 608 *n*, **651**
Kahneman, D., 164, 485, **651**
Kaiser, H. F., 83 *n*, **651**
Kanfer, F. H., 604, 605, **651**
Kaplan, L. J., 586 *n*, **651**
Karp, S. A., 591, **678**
Kass, E. H., 504, **645**
Kassorla, I. C., 608, **657**
Kastenbaum, R., 606, **630**

Katkin, E. S., 582, **671**
Katz, E., 284, **651**
Katz, I., **650**
Katz, M. R., 36, 95, 540, **623**, **651**
Katzell, R. A., 194, **653**
Kaufman, A. S., 260, 262, 263, 264, 275, 276, 428, 480, **647**, **651**, **652**
Kaufman, N. L., 275, 276, 428, 480, **651**, **652**
Kavruck, S., 17, **652**
Keller, H. R., 280, **637**
Kelley, T. L., 15, 16, 33, 208, 363, 372, 401, **652**
Kelly, E. L., 596, 597, **652**
Kelly, G. A., 598, 603, **652**
Kelman, H. C., 545, **652**
Kendall, L. M., 609, 610, **670**
Kendrick, S. A., 57, 192, 195, **631**, **652**
Keniston, K., 520, 521, **632**
Kennedy, W. A., 244, **652**
Kent, G. H., 570, **652**
Kent, R. N., 521, **641**
Keppel, G., 571, **665**
King, W. L., 335, **652**
Kinslinger, H. J., 581, **653**
Kinter, M., 454, **653**
Kirchner, W. K., 34, **653**
Kirk, G. E., 425, 431, **649**
Kirk, S. A., 479 *n*, 481, **653**
Kirk, W. D., 481, **653**
Kirkpatrick, J. J., 183 *n*, 194, **653**
Klaus, R. A., 285, **653**
Klausmeier, H. J., 38, **637**
Kleinmuntz, B., 487, 493, **653**
Klineberg, O., 288, **653**
Klinger, E., 578, 585, **653**
Knapp, R. R., 288, **653**
Knapp, S. C., 504, **661**
Knobloch, H., 267, 333, 334, 344, **653**, **664**
Knoell, D. M., 135, **653**
Koch, H. L., 351, **653**
Koch, S., **645**
Kogan, L., 58, **633**
Kogan, N., 604, **661**
Kohlberg, L., 554, 555, **653**
Koppitz, E. M., 475, **654**
Kovach, J. K., 5 *n*, **662**
Kraepelin, E., 10, 18, 569, **654**
Krasner, L., 523, **654**
Krathwohl, D. R., 136, 416, **654**
Krausz, M., 610, **679**
Krug, R. E., 297, **668**
Kuder, G. F., 117, 537, 538, 539, 540, 552, **654**
Kuhlen, R. G., 339, **654**
Kuhlmann, F., 12, **654**
Kunce, J. T., 541, **654**
Kurtines, W., 555, **654**
Kurtz, A. K., 220, **654**
Kuze, T., 34, **674**

Labouvie, G. V., 339, 340, **668**
Labouvie-Vief, G., 340, **668**
Lacey, J. I., 449, 562, 596, **644**
La Fave, L., 302, 522, 545, 593, **654**
La Forge, G. E., 169, **666**
Landfield, A. W., 604, **638**
Lang, P. J., 605, **679**
Lanyon, R. I., 493, 504, 519, **654**, **655**
Lao, R., 557, **644**
Laosa, L. M., 295, **655**
Larkin, K. C., 303, 304, **655**
Larson, G. R., 489 *n*, **670**
Laumann, E. O., 617, **655**
Laurendeau, M., 77, 277, 278, **655**, **664**
Lawrence, S. W., Jr., 38, **655**
Lawshe, C. H., 150, 435, 609, **655**
Layton, W. L., 505 *n*, **626**
Learned, J., 563, **621**
Lee, R., 615 *n*, **655**
Lehman, P. R., 458 *n*, **655**
Lehtinen, L. E., 467, **671**
Lejeune, S. J., 471
Lennon, R. T., 91, 135 *n*, **655**
Lens, W., 356, **624**
Lerner, J. W., 479 *n*, **655**
Lessler, K., 348, **648**
Levinson, B. M., 350, 467, **655**, **656**
Levitt, E. E., 563, **656**
Levonian, E., 509, **656**
Levy, P., 249, **656**
Lewinsohn, P. M., 606, **658**
Lewis, C., 118, **663**
Lewis, H. B., **678**
Lewis, M., 333, 335, 336, **656**
Lighthall, F. F., 37, 38, **667**, **676**
Likert, R., 550, 551, **656**
Lillesand, D. B., 606, **659**
Lindgren, B. W., 132, **656**
Lindquist, E. F., 319
Lindsley, D. B., 570, **656**
Lindzey, G., 351, 552, 558, 559, 567, 580, 613, **656**, **657**
Linn, R. L., 176, 196, 197, 303, **631**, **656**
Lipinski, D., 605 *n*, **656**

Littell, W. M., 246, 259, 260, 656
Little, K. B., 583, 656
Livingston, S. A., 132, 656
Locke, J., 8
Loehlin, J., 351, 657
Loevinger, J., 571, 598, 599, 657
Longstaff, H. P., 537, 657
Lonner, W. J., 297, 534, 628, 657
Lord, F. M., 205, 207, 208, 303, 304, 657
Loret, P. G., 92, 657
Loretan, J. O., 61, 77, 657
Lorge, I., 342, 550, 637, 657
Lotsof, E. J., 562, 657
Lovaas, O. I., 608, 657
Lowell, E. L., 568, 659
Luborsky, L. B., 592, 630, 657
Ludlow, H. G., 655
Lunneborg, C., 649
Lunzer, E. A., 277 *n*

MacDonald, M., 524, 657
Maceachern, D. G., 575, 676
Macfarlane, J. W., 329 *n*, 330, 648
Machover, K., 575, 658, 678
MacKinney, A. C., 148, 658
MacKinnon, D. W., 389, 595, 662
MacPhillamy, D. J., 606, 658
Macy, J. W., Jr., 61
Mahoney, M. H., 194, 436, 629
Maier, M. H., 189, 194, 196, 658
Mair, J. M. M., 603 *n*, 624
Maller, J. B., 595, 613, 645
Mallinson, G. G., 537, 658
Mancuso, J. C., 652
Mandler, G., 38, 658
Mann, H., 16
Marion, R., 196, 669
Marjoribanks, K., 618, 658, 676
Marker, R. W., 96, 638
Marks, P. A., 501, 658
Marlowe, D., 516, 632
Marschak, M., 559, 639
Marshall, N., 522, 545, 593, 654
Masling, J., 39, 40, 577, 578, 658
Masterson, S., 600 *n*, 658
Matarazzo, J. D., 246, 249, 252, 253, 465, 466, 467, 468, 470, 608 *n*, 658
Maxfield, K. B., 284, 659
Maxwell, E., 249, 659
May, M. A., 19, 594, 595, 613, 645
McArthur, D. L., 618, 645
McCall, R. B., 327, 329 *n*, 331, 333, 334, 335, 659
McCall, W. A., 83, 659
McCarthy, D., 40, 294, 659
McCarthy, J. F., 479 *n*, 659
McCarthy, J. J., 479 *n*, 659
McClelland, D. C., 568, 585, 659
McCormick, E. J., 150, 435, 439, 608 *n*, 612, 615 *n*, 659
McDonald, R. L., 504, 659
McElrath, G. W., 132, 656
McFall, R. M., 606, 659
McGurk, H., 335, 656
McKechnie, G. E., 616, 659
McKinley, J. C., 501, 646
McLeish, J., 457, 659
McLemore, S. D., 617, 627
McNamara, W. J., 597, 643
McNemar, Q., 224, 237, 240, 242, 243, 245, 249, 263, 659
McQuitty, J. V., 211, 662
McReynolds, P., 493, 626, 632, 634, 641, 648, 650, 658, 659, 660, 662, 675
Mecham, R. C., 439, 659
Meehl, P. E., 161, 172, 485, 486, 489 *n*, 496, 501, 632, 646, 660
Meeker, M., 617, 676
Meeker, M. N., 375, 660
Megargee, E. I., 493, 505, 515, 564, 567, 660
Meier, N. C., 453, 454, 660, 674
Meissner, P. B., 678
Melton, A. W., 443, 444, 660
Melville, S. D., 179, 639
Merrens, M. R., 489 *n*, 660
Merrill, M. A., 32, 230, 231, 232, 234, 237, 239, 241, 243, 244, 672, 673
Messick, S., 267, 336, 433, 520, 521, 589, 604, 619, 623, 626, 649, 660, 661
Metraux, R. W., 563, 621, 622
Miklich, D. R., 608, 661
Milgram, N. A., 285, 661
Miller, C., 504, 661
Miller, E. L., 339, 624
Miller, J. D., 92
Miller, L. P., 638
Miller, R. J., 347, 661
Millman, J., 44, 132, 661
Minium, E. W., 355, 665
Mischel, W., 493, 521, 522, 523, 545, 557, 604, 616, 618, 661
Mitchell, B. C., 195, 428, 661
Modu, C. C., 348, 623
Mollenkopf, W. G., 190, 217, 218, 661
Monachesi, E. D., 497, 501, 646

Moore, B. V., 608 *n*, 626
Moos, R. H., 616, 617, 661, 662
Moreno, J. L., 612, 662
Moriarty, A. E., 39, 235, 332, 662
Morrow, R. S., 454, 662
Moseley, E. C., 564, 648
Moses, J. L., 195, 627
Mosier, C. I., 211, 219, 662
Mowrer, O. H., 601, 662
Mugaas, H. D., 537, 662
Murphy, G., 5 *n*, 662
Murray, H. A., 510, 513, 565, 566, 595, 600, 619, 662
Murstein, B. I., 558, 568, 585, 662
Mussen, P. H., 583, 625, 662, 679
Myers, I. B., 518, 662
Myklebust, H. R., 479 *n*, 480, 650

Naylor, H. K., 583, 662
Naylor, J. C., 172 *n*, 662
Nebergall, R., 550, 669
Nelson, R. O., 604, 605 *n*, 636, 656
Nelson, V. L., 327, 329 *n*, 331, 650, 670
Nesselroade, J. R., 339, 648, 662
Newbert, N., 518, 669
Newman, S. H., 463, 663
Neyman, C. A., Jr., 91
Nichols, P. L., 351, 663
Nisbet, J. D., 339, 663
Nitko, A. J., 132, 133, 302, 305, 421, 641
Norman, R. D., 467, 663
Norman, W. T., 509, 664
North, R., 58, 633
Novick, M. R., 118, 132, 133, 197, 205, 208, 637, 645, 657, 663, 664
Nowicki, S., Jr., 557, 663
Nunnally, J. C., 415, 663

O'Connor, E. J., 194, 663
Odbert, H. S., 508, 621
Oden, M. H., 238, 324, 339, 625, 673
Oehrn, A., 10, 663
Oetting, E. R., 501, 634
Oleno, T., 610, 679
O'Malley, P. M., 356, 624
O'Neill, P. J., 615 *n*, 672
Oppel, W., 329 *n*, 677
Opper, S., 77, 640
O'Rourke, L. J., 17
Orr, D. B., 91
Ortar, G., 347, 663
Ortmeyer, D. H., 124 *n*, 646
Osborne, D., 502, 672
Osgood, C. E., 481, 601, 602, 603, 663, 670
Ossorio, A. G., 598, 599, 657
Otis, A. S., 12
Owens, W. A., 339, 342, 615, 663
Ozer, M. N., 285, 661

Palermo, D. S., 571, 663
Palmer, F. H., 329 *n*, 330, 665
Palmer, J. O., 32, 39, 663
Palmore, E., 341, 663
Paraskevopoulos, J., 279, 649, 663
Parker, G. V., 515, 567, 660
Pasamanick, B., 267, 333, 334, 344, 653, 664
Pascal, G. R., 473, 664
Passini, F. T., 509, 664
Paterson, D. G., 489 *n*
Patterson, C. H., 449, 541, 664
Patterson, G. R., 607, 608, 650
Paul, G. L., 38, 664
Pearson, J. S., 502, 664, 672
Pearson, K., 9, 363, 664
Peel, E. A., 44, 664
Peizer, E., 573, 664
Pemberton, C. L., 589, 664
Pennington, L. A., 648
Pepinsky, H. B., 597, 664
Petersen, N. S., 197, 664
Peterson, D., 522, 545, 664
Peterson, J., 5 *n*, 9, 664
Peterson, R. F., 604, 626
Philippe, J., 9, 664
Piaget, J., 77, 276, 277, 279, 432, 555
Pike, L. W., 43, 664
Pildner, H., Jr., 577, 633
Pinard, A., 77, 277, 278, 655, 664
Pinder, C. C., 180, 664
Pinneau, S. R., 127, 237, 240, 329 *n*, 334, 665
Piotrowski, Z. A., 563, 665
Pitcher, B., 460, 462, 665
Pittluck, P., 583, 665
Plant, W. T., 285, 355, 648, 665
Popham, W. J., 98 *n*, 132 *n*, 155, 665
Porter, A. C., 195, 671
Porteus, S. D., 288
Postman, L., 571, 665
Pressey, S. L., 342, 633
Price, L. A., 373, 639
Price, P. B., 149, 666
Prien, E. P., 148, 665
Primoff, E. S., 150, 151, 665

Quereshi, M. Y., 43, 665

Rabin, A. I., 246, 558, 647, 665
Radcliffe, J. A., 515, 665
Radloff, R., 615 *n*, 646
Raju, N. S., 169, 644
Ramsey, P. H., 245, 665
Rapaport, D., 465, 569, 665
Raven, J. C., 291, 292
Ray, E., 504, 629
Raynor, J. O., 624, 675
Redmore, C., 571, 598, 657
Rees, A. H., 329 *n*, 330, 665
Reese, H. W., 339, 648, 662
Reichenberg-Hackett, W., 40, 665
Reid, J. B., 607, 608, 650
Reid, J. W., 537, 665
Reilly, R. R., 196, 666
Reinehr, R. C., 564, 648
Reitan, R. M., 468, 666
Reynolds, H. H., 613, 666
Rich, C. C., 284, 666
Richards, J. M., Jr., 149, 666
Richards, W. S., 489 *n*, 660
Richardson, M. W., 117, 172, 654, 666
Rider, R. V., 329 *n*, 677
Rimoldi, H. J., 291, 666
Rizzo, G. B., 504, 667
Roback, H. B., 575, 666
Robb, G. P., 488, 666
Robertson, M. H., 578, 645
Robinson, D. D., 615 *n*, 666
Robinson, J. P., 552, 666
Robinson, N. M., 329 *n*, 330, 627
Rock, D. A., 176, 194, 303, 436, 629, 631, 656
Rodell, J. L., 563, 621
Rodger, A. G., 44, 666
Roe, A., 542, 666
Rogers, C. R., 601, 666
Rohila, P. K., 542, 674
Rorer, L. G., 169, 485, 520 *n*, 521, 582, 641, 666
Rorschach, H., 560, 666
Rosanoff, A. J., 570, 652
Rosen, A., 172, 660
Rosen, E., 504, 667
Rosenberg, L. A., 285, 667
Rosenberg, N., 537, 667
Rosenstein, A. J., 523, 636
Rosenthal, R., 39, 40, 667
Rosenthal, T. L., 281, 679
Rosenzweig, S., 572
Rosnow, R. L., 40, 667
Ross, D. R., 622
Ross, J., 649
Rothney, J., 43, 633
Rotter, J. B., 523, 555, 556, 667
Royce, J. R., 661
Rubin, S. E., 606, 672
Ruch, G. M., 16
Ruch, W. W., 196, 667
Ruebhausen, O. M., 51, 53, 54, 667
Ruebush, B. K., 37, 667
Rulon, P. J., 116, 667
Russell, J. T., 170, 171, 172, 672
Russell, W. A., 571, 650

Sachman, H., 562, 667
Sacks, E. L., 34, 667
Samelson, F., 521, 667
Sanders, E. P., 179, 640
Sanford, N., 599, 667
Sarason, I. G., 37, 493, 667
Sarason, S. B., 37, 38, 39, 469, 578, 647, 658, 667, 676
Sarbin, T. R., 483, 484, 668
Saslow, G., 604, 651
Sattler, J. M., 32, 39, 230 *n*, 234, 243, 245, 263, 264, 281, 470, 471, 480, 488, 668
Saunders, D. R., 178, 183 *n*, 438, 639, 668
Sawyer, J., 485, 668
Schaefer, C. E., 615 *n*, 623
Schaefer, E. S., 329, 625
Schaie, K. W., 339, 340, 341, 668
Schickedanz, D., 279, 649
Schmidt, F. L., 194, 197, 668
Schulberg, H. C., 473, 474, 674
Schumer, F., 558, 679
Schutz, R. E., 379, 668
Schwartz, R. D., 607, 676
Schwarz, P. A., 297, 668
Scott, W. A., 519, 668
Scott, W. E., Jr., 541, 668
Seashore, C. E., 455, 456, 669
Seashore, H. G., 178, 255, 259, 263, 626, 669
Seashore, S. E., 149, 669
Sechrest, L., 171, 607, 669, 676
Seder, A., 92, 657
Seegmiller, B., 335, 652
Seeman, W., 501, 658
Segall, M. H., 347, 669
Seguin, E., 6, 669

Sewell, W. H., 624
Shanner, W. M., 96, 638
Sharp, S. E., 9, 669
Shaver, P. R., 552, 666
Shaw, M. E., 551, 669
Shaycoft, M. F., 91
Shepps, R. R., 615 *n*, 672
Sherif, C. W., 550, 669
Sherif, M., 550, 648, 669
Shimberg, B., 36, 623
Shimberg, M. E., 222, 669
Shine, L. C., 172 *n*, 662
Shipley, W. C., 518, 669
Shneidman, E. S., 558, 583, 656, 669
Shore, C. W., 196, 669
Shulman, A. D., 516, 669
Shuttleworth, F. K., 645
Siegel, L., 597, 664
Siess, T. F., 541, 669
Sigel, I. E., 302, 329, 336, 669
Silverman, I., 516, 669
Silverman, L. H., 581, 583, 669
Silverstein, A. B., 249, 257, 260, 262, 669, 670
Simon, Th., 11, 27, 626
Sisson, D. E., 518, 670
Sivers, C. H., 348, 647
Skolnick, A., 585, 670
Slingerland, B. H., 480
Sloan, W., 269, 670
Smith, A., 467, 670
Smith, G. H., 573, 670
Smith, L. F., 455, 648
Smith, P. C., 609, 610, 670
Smith, R. S., 334, 677
Smith, S., 346, 670
Smith, T. A., 348, 629
Smith, W. J., 149, 621
Snider, J. G., 603, 670
Snyder, C. R., 489 *n*, 670
Snygg, D., 598, 670
Sommer, R., 18, 670
Sontag, L. W., 327, 329 *n*, 331, 650, 670
Sorenson, W. W., 183, 670
Spearman, C., 14, 363, 369, 370, 371, 670
Spielberger, C. D., 37, 640, 641, 649, 670
Sprafkin, J. N., 604, 641
Spranger, E., 552, 553, 670
Spuhler, J. N., 351, 657
Stachnik, T., 489 *n*, 674
Stagner, R., 489 *n*, 671
Stainton, R., 489 *n*, 674
Stanley, J. C., 195, 329, 336, 415, 617, 635, 671
Starr, B. J., 582, 671
Starr, R. P., 410, 675
Stauffer, G. F., 169, 633
Steckler, V., 387, 671
Stephens, M. W., 557, 671
Stephenson, W., 601, 671
Stern, G. G., 619, 671
Sterritt, G. M., 649
Stott, L. H., 245, 272 *n*, 334, 671
Strauss, A. A., 467, 469, 671, 677
Stricker, G., 562, 641
Stricker, L. J., 179, 515, 671
Strong, D. J., 599, 671
Strong, E. K., Jr., 529, 543
Strother, C. R., 339, 341, 668
Stroud, M., 285, 667
Strunk, W., Jr., 489 *n*, 671
Su, W. H., 197, 643
Subkoviak, M. J., 661
Suci, G. J., 663
Summers, G. F., 545 *n*, 551, 671
Sundberg, N. D., 489 *n*, 504, 542, 671, 674
Super, D. E., 95, 334, 378, 383, 435, 542, 554, 671, 672
Suttell, B. J., 473, 664
Swartz, J. D., 295, 563, 648, 655
Swensen, C. H., 575, 672
Swenson, W. M., 502, 664, 672
Swets, J. A., 478 *n*, 672
Symonds, P. M., 18, 672

Taft, R., 483, 504, 668, 672
Talmage, H., 638
Tannenbaum, P. H., 663
Tanofsky, R., 615 *n*, 672
Tasto, D. L., 606, 672
Tatsuoka, M. M., 509, 630
Taylor, C. W., 149, 389, 615 *n*, 650, 666, 672
Taylor, H. C., 170, 171, 172, 672
Taylor, J. A., 505 *n*, 672
Teahan, J. E., 348, 672
Tellegen, A., 249, 672
Temp, G., 196, 672
Terman, L. M., 12, 16, 32, 230, 231, 232, 234, 237, 238, 239, 241, 243, 244, 324, 672, 673
Tharp, R. G., 606, 673
Theologus, G. C., 439, 673

Theye, F., 39, 668
Thomas, C. L., 195, 652
Thomas, H., 272 *n*, 334, 673
Thompson, C. W., 327, 627
Thomson, G. H., 376, 673
Thorndike, E. L., 16
Thorndike, R. L., 190, 197, 311, 312, 313, 328 401 *n*, 623, 624, 631, 641, 646, 673, 677
Thorndike, R. M., 297, 628
Thorpe, J. S., 563, 648
Thurstone, L. L., 15, 363, 365, 372, 378, 389, 447, 546, 547, 548, 549, 550, 551, 589, 673, 674
Thurstone, T. G., 372, 674
Thwing, E., 334, 649
Tiebout, C., 454, 674
Tiffin, J., 435, 608 *n*, 612, 615 *n*, 659
Todd, F. J., 485, 645
Tolor, A., 473, 474, 674
Torrance, E. P., 389, 394, 396, 674
Tracht, V. S., 284, 674
Traxler, A. E., 33, 674
Trumbo, D., 608 *n*, 674
Truumaa, A., 563, 656
Tryon, R. C., 376, 674
Tsudzuki, A., 34, 674
Tucker, L. R., 523, 674
Tuddenham, R. D., 277 *n*, 339, 353, 674
Turner, J. N., 194, 625
Tversky, A., 485, 651
Twentyman, C. T., 606, 659
Tyler, L. E., 542, 543, 674

Uçman, P., 295, 674
Uhrbrock, R. S., 549, 674
Ulrich, L., 608 *n*, 674
Ulrich, R., 489 *n*, 674
Upshaw, H. S., 550, 675
Užgiris, I. Č., 77, 278, 279, 649, 675

Vale, C. A., 92, 657
Valett, R. E., 482
Vanatta, A., 597, 664
Van de Reit, V., 244, 652
Van der Made-Van Bekkum, I. J., 571, 675
Vane, J. R., 245, 665
Varble, D. L., 565, 675
Velicer, W. F., 180, 675
Vernon, P. E., 297, 347, 375, 376, 552, 597, 675
Veroff, J., 567, 568, 675
Vinh-Bang, N. P., 277 *n*
Vinitsky, M., 534, 675
Vitola, B. M., 388, 675

Waggener, R., 410, 675
Wahlstrom, M., 44, 675
Wainwright, W. H., 483, 676
Waite, R. R., 37, 38, 667, 676
Walberg, H. J., 618, 676
Wald, A., 132, 169, 676
Walker, D. K., 21, 676
Walker, R. N., 563, 621, 622
Wallace, J. G., 44, 646
Wallace, S. R., 149, 676
Wallace, W. L., 250, 318, 337, 634, 676
Wallach, M. A., 389, 676
Walsh, T. M., 593, 676
Walsh, W. B., 616, 676
Walters, R. H., 521, 523, 624
Wand, B., 438, 639
Wapner, S., 678
Warburton, F., 589, 630
Warner, W. L., 617, 676
Warrington, W. G., 124 *n*, 205, 632
Wasik, B. H., 280, 676
Wasik, J. L., 281, 676
Watson, C. G., 575, 676
Watson, J. B., 523, 676
Watson, R. R., 523, 676
Waugh, K. W., 479 *n*, 480, 676
Webb, E. J., 607, 608, 676
Wechsler, D., 245, 246, 252, 253, 256, 262, 465, 466, 470, 676
Weiner, I. R., 562, 641
Weiss, D. J., 176, 302, 303, 304, 633, 655, 677
Weissenberg, P., 591, 677
Weitz, J., 613, 677
Welsh, G. S., 499, 500, 501, 502, 503, 592, 600, 633, 677
Wender, P. H., 479 *n*, 677
Werner, E. E., 334, 677
Werner, H., 467, 677
Wertheimer, M., 473
Weschler, I. R., 559, 677
Wesman, A. G., 57, 192, 218, 255, 259, 263, 415, 516, 626, 631, 669, 677
Wessler, R., 571, 598, 657
Wetzel, R. J., 606, 673
Wexley, K. N., 194, 663
Wheeler, L. R., 353, 677
Whimbey, A. E., 329, 377, 677
White, E. B., 489 *n*, 671

White, J. C., 244, 652
Whiteman, M., 58, 377, 633, 677
Wickes, T. A., Jr., 34, 677
Wiener, G., 329 *n*, 677
Wiens, A. N., 608 *n*, 658
Wiggins, J. S., 172, 175 *n*, 177 *n*, 183, 485, 486 *n*, 493, 499, 515, 520, 595, 596, 607, 608, 613, 677, 678
Wiggins, N., 485, 519, 678
Wiley, D. E., 660
Wilkerson, D. A., 329, 642
Wilkin, W. R., 339, 674
Wilks, S. S., 196, 644
Willard, L. S., 291, 348, 678
Williams, M., 342, 678
Willingham, W. W., 322, 678
Willis, J., 302, 678
Winder, C. L., 613, 678
Wing, C. W., Jr., 389, 676
Wing, H. D., 457, 678
Wissler, C., 9, 10, 678
Witkin, H. A., 589, 591, 678
Wittrock, M. C., 660
Wohlwill, J. F., 279 *n*, 678
Wolf, R., 618, 678
Wolf, T. H., 6, 11, 678
Wolk, R. B., 569, 679
Wolk, R. L., 569, 679
Wolman, B. B., 645, 649, 669
Wolpe, J., 523, 605, 679
Womer, F. B., 52, 97, 205, 679
Wonderlic, E. F., 441
Woodworth, R. S., 18, 494
Woo-Sam, J. M., 246, 259, 260, 466, 467, 470, 488, 679
Wright, H. F., 607, 608, 679
Wright, J. M., 551, 669
Wrightstone, J. W., 403, 679
Wundt, W., 7, 8, 9, 569

Yamamoto, K., 396, 679
Yates, A. J., 42, 679
Yerkes, R. M., 12, 265, 679

Zedeck, S., 610, 645, 679
Zeiss, A., 557, 661
Zeiss, R., 557, 661
Zimbardo, P., 37, 667
Zimmerman, B. J., 281, 679
Zimmerman, I. L., 246, 259, 260, 466, 467, 470, 488, 641, 679
Zimmerman, W. S., 506, 644
Zubek, J., 339, 625
Zubin, J., 558, 679

SUBJECT INDEX

Academic Promise Tests (APT), 711
Achievement motivation (n-Ach), 356–357, 568, 573–574, 585
Achievement tests, 16–17, 398–423, 711–716
  for adult basic education, 408–410, 711–712
  diagnostic, 417–423, 713–714
  for early childhood education, 425–434, 714
  general batteries, 403–410, 711–712
  nature of, 398–402
  origins of, 16–17
  in professions, 458–463
  relation to aptitude tests, 398–402
  in separate subjects, 410–412, 712–713
  teacher-made, 412–417
  uses of, 402–403
  vocational, 17, 439, 715–716
Acquiescence, as response set, 520
ACT Battery of American College Testing Program, 319–320, 710
ACT Career Planning Profile, 711
Activities Index, 619, 721
Adaptive Behavior Scale (ABS), 270–271, 709
Adaptive testing, 302–305
Adjective Check List (ACL), 599–600, 613–614, 720
Adult Basic Learning Examination (ABLE), 408–410, 711
Adult testing, 36–37, 245–255, 313, 318–325, 337–343, 408–410, 435–463
Age decrement in intelligence test scores, 337–338
Age norms, 74–75
Allport–Vernon–Lindzey Study of Values, 552–554, 719
Alpha Examination, Modified Form 9, 299, 710
Alternate-form reliability, 112–114, 120–121
American Association on Mental Deficiency, Adaptive Behavior Scale, 270–271, 709
  levels of mental retardation, 238–239
American Board of Professional Psychology, 47, 463
American College Testing Program, 17
  ACT Battery, 319–320, 710
  ACT Career Planning Profile, 711
American Home Scale, 617, 721
American Psychological Association, ethical practices pertaining to tests, 45–57, 682–690
  recommendations on test construction and use, 22, 134, 140, 159, 192, 670
Analysis of Learning Potential, 306–307, 710
Analysis of Readiness Skills, 714
Anchor test, 91–92, 93, 411–412
Anchor Test Study, 92
Animal Crackers, 573–574, 719
Anonymity, and protection of privacy, 51, 55
Answer sheet, effect on test scores, 33–34
AO Sight Screener, 715
Appraisal of Occupational Aptitudes, 715
Aptitude Research Project (ARP), 390–394
Aptitude tests, early development, 13–16
  examples, 442–458, 715–716
  relation to achievement tests, 398–402
  (*see also* Special aptitude tests)
Arithmetic tests, 420–421
Armed Forces Qualification Test (AFQT), 189, 299, 301
Armed Services Vocational Aptitude Battery, 388
Army Alpha, 12–13, 299
Army Beta, 12–13, 281, 299, 710
Arthur Adaptation of the Leiter International Performance Scale, 289, 709
Arthur Point Scale of Performance Tests, 281, 709
Artistic aptitudes, tests of, 451–455, 716

Attitude scales, 544, 546–552, 719
Attitude shifting, 522
Attitudes, measurement of, 543–552
    applications, 544
    methodological problems, 545–546
    projective techniques, 559, 568, 576
    types of scales, 546–552
Auditory Discrimination Test, 717
Auditory tests, 457, 715, 716, 717
Average deviation, 72

Barnum effect, 489, 581 *n*
Barron–Welsh Art Scale, 591–592, 720
Basal age, 74, 235–236
Base rate, 170–172
Basic Occupational Literacy Test (BOLT), 387, 410, 712
Battery, 180–181
Bayley Scales of Infant Development, 271–273, 709
Behavior modification, assessment techniques for, 604–606, 616
    effect on personality test construction, 522–524
Behaviorally anchored rating scales, 609–610
Bender–Gestalt Test, 473–475, 480, 717
    Koppitz standardization, 475
    Pascal–Suttell standardization, 473–475
Bender Visual Motor Gestalt Test (*see* Bender–Gestalt Test)
Bennett Mechanical Comprehension Test, 448–449, 715
Benton Visual Retention Test, 471–473, 480, 717
Bibliographies of tests, 20–21
Binet scales (*see* Binet–Simon Scales)
Binet–Simon Scales, 10–12
    revisions, 12, 74, 230
Biographical inventories, 614–615
Biserial correlation, 214–215
Bivariate distribution, 104–106, 164–165, 192–196, 222–223
    heteroscedastic, 164–165
Blacky Pictures, 720
Blind, tests for, 282–284
Blyth Second-Year Algebra Test—Revised Edition, 712
Boehm Test of Basic Concepts, 429–431, 714
Brain damage, diagnosis, 467–478
Buswell–John Diagnostic Test for Fundamental Processes in Arithmetic, 420–421, 713

C scale, 83 *n*
Cain–Levine Social Competency Scale, 709
California Achievement Tests, 403, 404, 712
California Occupational Preference Survey, 719
California Psychological Inventory (CPI), 505–506, 718
California Q-set (CQ-set), 614
California Test of Mental Maturity, 306–307, 710
Capacity, concept of, 25
Career Maturity Inventory, 719
Categorizing tendency, in use of test results, 60–61
Cattell Infant Intelligence Scale, 709
Ceiling age, 235–236
Central tendency, 70–72
Cerebral palsied, tests for, 284–286
Certification of psychologists, 47
Chapin Social Insight Scale, 720
Character Education Inquiry (CEI), 594–595
Children's Apperception Test (CAT), 568–569, 720
Children's Personality Questionnaire, 509, 718
Circus, 432–434, 480, 618–619, 714
Civil service examinations, 5, 17, 61–62
Classification, use of tests for, 186–191
Classroom tests, 209–211
Clerical aptitude tests, 449–451, 715–716
Clinical Analysis Questionnaire, 510, 718
Clinical judgment, 464, 482–487, 587
Clinical use of tests, 3, 10, 464–490
    (*see also* Intelligence tests, individual; Personality testing)
Clinical versus statistical prediction, 485–487
Closure Flexibility (Concealed Figures), 589–590, 720
Closure Speed (Gestalt Completion), 589–590, 720
Coaching, and test performance, 41–43
Coefficient alpha, 118
Cognitive Abilities Test (CAT), 306, 310–313, 710

Cognitive dysfunction, tests for, 471–478, 717–718
College Characteristics Index (CCI), 619–620, 721
College Entrance Examination Board, achievement tests, 411–412
  Advanced Placement Program (APP), 412
  College-Level Examination Program (CLEP), 412
  effects of coaching, 42–43
  establishment of testing program, 17
  Preliminary SAT, 318
  Scholastic Aptitude Test (SAT), 81–82, 87, 93–94, 283, 301, 318–319, 348–349, 710
  score scale, 87, 93–94, 411–412
  test orientation booklets, 36
  test performance of minority groups, 195–196, 222–224
College environments, measurement of, 617–618, 619–620
College Qualification Tests, 320, 710
College and University Environment Scales (CUES), 620, 721
Columbia Mental Maturity Scale, 285–286, 709
Columbian Exposition, 9
Combining tests, methods for, 180–186
Communicating test results, 56–57
Complex Coordination Test, 446
Computer-assisted instruction (CAI), 96
Computer-managed instruction (CMI), 96
Computers, and confidentiality of records, 54–55
  in group testing, 302–305
  in interpretation of test scores, 94–96
    CPI, 506
    DAT, 95, 382–383, 486–487
    Holtzman Inkblot Technique, 564
    MMPI, 95, 486–487, 502–503
    Rorschach, 563
  in scoring interest inventories, 529, 538
Comrey Personality Scales, 507–508, 718
Concept Assessment Kit—Conservation, 279–281, 709
Concept Formation Test, 717
Concept Mastery Test, 324–325, 710
Concurrent validity, 140–141
Confidentiality, 52–55
Conflict Resolution Inventory, 606
Construct validity, 151–158, 438
  relation to other kinds of validity, 158–161
Contemporary Mathematics Test, 712
Content Evaluation Series, 712
Content meaning, in interpretation of test scores, 97–98, 100
Content validity, 134–140
Convergent validation, 156–158
Cooperative Mathematics Tests, 712
Cooperative Preschool Inventory, 714
Cooperative Primary Tests, 712
Cooperative Science Tests, 713
Correlation, 104–110
  biserial, 214–215
  dependence upon range, 125–127, 163–165
  heteroscedasticity, 165
  multiple, 182
  nonlinear, 164
  Pearson Product-Moment, 106–108
  phi, 213–214
  significance of, 108–109, 165
Counseling, use of tests in, 4, 56–57, 435
  (*see also* Interest tests; Multiple aptitude batteries; Occupational testing)
Crawford Small Parts Dexterity Test, 444, 715
Creative achievement, 396–397
Creativity tests, 388–397, 711
Creativity Tests for Children, 711
Criterion, in test validation, 28–29, 142–149
    specificity, 146–149
Criterion contamination, 141–142
Criterion keying, of biographical inventory items, 615
  of interest inventory items, 529, 533, 534, 538–539, 540
  of personality inventories, 496–506
Criterion-referenced testing, 96–102, 131–133, 138, 422–423, 713–714
Criterion-related validity, 140–151, 436–437
Critical incident technique, 438
Critical part-activities, in job analysis, 437
Cross-cultural testing, 58–61, 191–197, 222–224, 287–298, 343–349
  approaches to, 295–298
Cross-sectional versus longitudinal studies, adult intelligence, 338–341

Cross validation, 182, 219–221
Cultural differentials, and interpretation of test scores, 59–61, 343–346
  levels of, 343–344
  in test performance, 58–59
Culture Fair Intelligence Test (IPAT), 289–291, 710
Culture-fair tests, 191–197, 222–224, 287–298, 343–349
Culture-free tests (*see* Cross-cultural testing)

D 48 Test, 710
Davis–Eells Games, 225
Deaf, tests for, 281–282
Decision theory, in test use and evaluation, 167–177, 186–190, 197
  adaptive treatments, 177
  sequential strategies, 176–177
  values, 174–175
Dental Admission Testing Program, 716
Derived scores, 67
Deterioration index on WAIS, 465
Developed abilities, concept of, 399–400
Developmental norms, 73–77
Developmental scales, 267–271, 284, 428, 709–710
Deviation, as response style, 520
Deviation IQ, 84–86, 236–239
Diagnosis: An Instructional Aid Series, 713
Diagnosis versus prediction, 24–25, 140–141
Diagnostic Reading Scales, 713
Diagnostic Reading Tests, 713
Diagnostic tests, in education, 417–423, 713–714
  mathematical skills, 420–422, 713–714
  reading, 417–420, 713–714
Differential aptitude testing, 13–16, 361–362
Differential Aptitude Tests (DAT), 101, 130–131, 147–149, 379–383, 711
  Career Planning Program, 382–383
Differential validity, 190, 383
Difficulty, measurement of, 26–27
  of test items, 199–206
Discriminant validation, 156–158
Distribution, bivariate, 104–106, 126, 164–165, 222–223
  frequency, 68–70, 201–204
  normal, 69–70, 73, 78–80
  rectangular, 203 *n*
  skewed, 202–203
Divergent production tests, 390–394
Draw-a-Man Test, 40, 293–295
Drawing, aptitude tests for, 454–455
  in intelligence testing, 293–295
  in personality assessment, 575–576
Drives, relation to intelligence, 355
Duplicating technique, 594
Durrell Analysis of Reading Difficulty, New Edition, 419–420, 713
Durrell Listening-Reading Series, 713
Dvorine Color Vision Test, 715

Early childhood education, tests for, 425–434, 714
Early School Personality Questionnaire, 509, 718
Ecological psychology, 616
Educational Testing Service (ETS), 17, 201
Educational tests, 398–434, 711–714
Educational use of tests, 3, 16–17, 96–102, 398–434
Edwards Personal Preference Schedule (EPPS), 510–513, 518, 519, 718
Ego development (I-level), 599
Elementary school children, group intelligence tests for, 310–313
Embedded Figures Test, 591, 720
Environment, assessment of, 616–620, 721
Environmental psychology, 616–617
Environmental Response Inventory, 616, 719
Equal-appearing intervals, method of, 547
Equal Employment Opportunity Commission (EEOC), 62–64, 192, 691–699
Equal-unit scaling, 82 *n*
Equipercentile method, 91
Error of central tendency, in rating, 611
Error-choice technique, 593
Error of measurement, 103, 127–131
Error variance, 103, 119–122
Essay questions, 414–415
Ethical Standards of Psychologists, 45–53, 682–690
Examiner, preparation, 32–33
  qualifications, 31
  reliability, 119
  variables, 39–40

  in projective techniques, 40, 562, 578
Examining for Aphasia, 717
Expectancy chart, 101–102, 163
Expectancy tables, 100–102, 163
Experience-producing drives, 355
Experimental psychology, influence on early test development, 7
Eysenck Personality Inventory, 718

Face validity, 139–140
Factor analysis, 14–15, 153–154, 362–369
  creativity, 390–394
  intelligence, 369–378
  motor functions, 445–446
  perception, 527–528
  personality inventories, 506–510
  reference tests, 373
  Stanford-Binet, 244–245
  Wechsler scales, 254–255, 260, 264
Factor matrix, 362–364
Factorial validity, 154, 366
Faking, on personality inventories, 498–499, 505, 513, 515–520
  on projective techniques, 577
False positives, 167
Famous Sayings test, 592–593, 721
Fear Survey Schedule, 605, 721
Field dependence, tests of, 591
Fixed reference group, in nonnormative scales, 93–94
Flanagan Aptitude Classification Tests (FACT), 711
Flanagan Industrial Tests (FIT), 711
Forced-choice technique, 517–520, 536–537, 553, 556
Forced distribution, 601, 611
Free association tests, 18, 569–571
Frequency distribution (*see* Distribution, frequency)
Frequency polygon, 69–70
Frostig Developmental Test of Visual Perception, 482, 717
Frostig Movement Skills Test Battery, 717
Full Range Picture Vocabulary Test, 709
Fundamental Achievement Series, 712

g factor, 369–370
Galton laboratory, influence on early test development, 8
Gates–McKillop Reading Diagnostic Tests, 713
General Aptitude Test Battery (GATB), 184–185, 384–387, 451, 711
General Clerical Test, 716
Gerontological Apperception Test, 569, 720
Gesell Developmental Schedule, 76, 267–268, 709
Gesell Developmental Tests, 428, 714
Gesell Institute, developmental scale, 76, 267–268, 709
  school readiness tests, 428, 714
Gilmore Oral Reading Test, 713
Goldman–Fristoe–Woodcock Test of Auditory Discrimination, 717
Goldstein-Scheerer Tests of Abstract and Concrete Thinking, 476–478, 717
Goodenough Draw-a-Man Test, 293–295
Goodenough–Harris Drawing Test, 293–295, 710
Gordon Occupational Check List, 719
Gordon Personal Inventory, 518, 718
Gordon Personal Profile, 718
Gottschaldt Figures, 589–590
Grade norms, 75–76
Graduate Record Examinations (GRE), 283, 321–323, 710
Graduate school students, tests for, 321–324
Graves Design Judgment Test, 716
Gray Oral Reading Test, 713
Group factors, 371–378
  interpretation, 366–368
Group tests, nature of, 299–305
  origin of, 12–13
  survey of, 305–325
Guess Who technique, 612–613
Guidelines on Employee Selection Procedures (EEOC), 62–64, 192, 691–699
Guidelines for Reporting Criterion-Related and Content Validity (OFCC), 63, 699–704
Guilford–Zimmerman Aptitude Survey, 711
Guilford–Zimmerman Interest Inventory, 719
Guilford–Zimmerman Temperament Survey, 506–507, 718
Guttman scale, 76 *n*, 98 *n*

Hall Occupational Orientation Inventory, 719
Halo effect in ratings, 611

Haptic Intelligence Scale for Adult Blind, 283, 709
Head Start, 329, 335–336
Henmon-Nelson Tests of Mental Ability, 1973 Revision, 306, 710
Heritability, of intelligence, 351–354
Heteroscedasticity, 165
Hierarchical theories of abilities, 375–376
High school students, group intelligence tests for, 313–318
Hiskey-Nebraska Test of Learning Aptitude, 281–282, 709
Histogram, 69–70
History of psychological testing, 5–19
Holtzman Inkblot Technique, 563–565, 720
Home Index, 617
Homogeneity of test items, 116–118, 154–155, 215–217
Honesty, tests for, 594
Horn Art Aptitude Inventory, 454–455, 716
H-T-P: House-Tree-Person Projective Technique, 720

Illinois Test of Psycholinguistic Abilities (ITPA), 481–482, 717
Illusory validation, 582
Improbable achievement tests, 594
In-Basket Test, 438–439
Incremental validity, 171–172
Index of discrimination (D), 211–213
Index of reliability, 120–121
Individual differences, in early experimental psychology, 7
  and early test development, 8–9
Individual Pupil Monitoring System, 713
Individual tests, 12, 229–264, 267–286, 288–289, 709–710
  comparison with group tests, 299–305
Industrial use of tests, 3–4, 17, 435–458
  with cultural minorities, 57–64, 191–197, 691–704
Infant testing, 266–273, 276–281, 332–335
Informed consent, 51–52
Inkblots, as projective tests, 559–565
Instability of IQ, reasons for, 329–332
Intelligence, composition of, 369–378
  and environmental variables, 329–331
  longitudinal studies, 327–335, 339–343
  modifiability, 350–354
  nature of, 334–335, 341–343, 349–357
  population changes in, 353–354
  relation to personality, 331–332, 354–357
Intelligence quotient (IQ), 12, 60–61, 84–86, 236–239, 327–332, 349–350
Intelligence tests, 10–13, 229–357
  classified list, 709–710
  diagnostic clinical use, 465–471
  group, 289–295, 299–305, 710
  individual, 229–264, 267–286, 288–289, 709–710
  for industrial screening, 439–442
  interpretation, 59–61, 343–346, 349–357
  longitudinal studies, 327–335, 339–341
  names for, 306
  qualitative observations, 470–471
Interactive computer systems, 95–96
Intercept bias, 195–197
Interest tests, 527–542, 593, 719
  methodological problems, 539–543
  and personality theory, 541–543
  relation to other personality measures, 541–543
Interim Evaluation Tests, 422, 713
Interim Hayes-Binet, 282–283
Interitem consistency, measures of, 116–118
Internal consistency, 154–155
  and item validity, 215–217
Internal–external locus of control (I-E scale), 555–557
Interpersonal perception, 483
Intervention programs, 329, 335–336
Interviewing techniques, 608–609
Intra-individual differences, significance, 129–131, 315–316, 465–467
Inventory of College Activities, 618, 721
Iowa Silent Reading Tests, 713
Iowa Tests of Basic Skills, 403, 404, 712
Iowa Tests of Educational Development, 404, 712
IPAR, 389, 600
IPAT Humor Test of Personality, 592, 721
Ipsative scores, 512–513, 518, 539–540, 553, 601
IQ (*see* Intelligence quotient)

Item analysis, 198–226
  and cross validation, 219–221
  and internal consistency, 215–217
  item × group interaction, 222–226
  in mastery tests, 206
  of speeded tests, 217–219
  of Stanford–Binet, 232
  techniques for, 199–214
  use of extreme groups, 208–214
Item characteristic curves, 207, 208
Item difficulty, 199–206
  group differences, 222–226
  and interval scales, 200–201
  and maximum differentiation, 203–205
  relation to D, 212–213
  relation to phi, 213
  and screening, 205
  in speed tests, 217–219
Item sampling, and test reliability, 112–114
Item validity, 206–215
  cross validation, 219–221
  and internal consistency, 215–217
  in speed tests, 217–219
Item writing, 413–417

J-coefficient, 150–151
Job analysis, 437–438
Job-sample tests, 439, 451
Jr.–Sr. High School Personality Questionnaire, 509, 718

Kent–Rosanoff Free Association Test, 570–571, 720
Keystone Telebinocular, 715
Kuder General Interest Survey (KGIS), 538, 719
Kuder interest inventories, 536–539, 719
Kuder Occupational Interest Survey (KOIS), 538–539, 719
Kuder Personal Preference Record, 719
Kuder–Richardson reliability, 116–118, 120, 123
Kuder Vocational Preference Record, 536–538, 719
Kuhlmann–Anderson Tests, 306, 710
Kuhlmann–Binet, 12

Lambda coefficient, 538 *n*
Landscape Adjective Check List, 616
Language, in transcultural testing, 346–349
Language Modalities Test for Aphasia, 717
Law School Admission Test, 460–462, 716
Leaderless Group Discussion (LGD), 597
Learning disabilities, 478–482
  assessment techniques, 479–482, 717–718
  nature, 478–479
Lee–Clark Reading Readiness Test, 714
Leiter International Performance Scale, 284, 288–289, 709
Leniency error in rating, 611
Let's Look at Children, 432, 714
Licensing of psychologists, 47
Lie detector, 570
Life history, 483, 609, 614–615
Likert-type attitude scales, 550–551
Lincoln–Oseretsky Motor Development Scale, 269, 481, 709
Link system, for protecting confidentiality, 55
Living Room Check List, 617
Logical Reasoning test, 397, 711
Longitudinal studies, of individual intelligence, 327–335, 339–341
  of populations, 353–354

Machover Draw-a-Person Test (D-A-P), 575–576, 720
MacQuarrie Test for Mechanical Ability, 715
Make a Picture Story (MAPS), 720
Manual, information to be included in, 21
Massachussetts Hearing Test, 715
Mastery testing, 98–100, 131–133, 206
Mathematical skills, diagnostic tests, 420–422, 713–714
  prognostic tests, 423, 714
  readiness tests, 426–429, 714
McCarthy Scales of Children's Abilities, 273–276, 480, 709
McGraw-Hill Basic Skills System, 410, 712
Mean, computation of, 70–71
Mechanical aptitude tests, 446–449, 715
Median, 70–72
Mediator–Reinforcer Incomplete Blank, 606
Medical College Admission Test, 716
Meier Aesthetic Perception Test, 716
Meier Art Judgment Test, 452–454, 716
Memory-for-Designs Test, 717

Mental age, concept of, 74–75
  introduction of, 11
  Stanford–Binet, 235–237
Mental level, 11
Mental Measurements Yearbooks, 20–21, 229, 708
Mental retardation, use of tests in identification, 3, 5–7, 238–239, 269–271
"Mental test," early examples, 9–10
  first use of term, 9
Metropolitan Achievement Test, 404, 712
Metropolitan Readiness Tests, 426–428, 480, 714
Military use of tests, 4, 12–13, 15, 18, 19, 40, 187, 189, 299, 387–388, 444, 595–596
Miller Analogies Test, 323–324, 710
Minnesota Clerical Test, 449–451, 716
Minnesota Counseling Inventory, 505 *n*, 718
Minnesota Engineering Analogies Test (MEAT), 716
Minnesota Multiphasic Personality Inventory (MMPI), 497–504, 718
  evaluation, 502–504
  profile coding, 501–503
  response styles, 521
  status scale, 225
Minnesota Paper Form Board, 447–448, 715
Minnesota Teacher Attitude Inventory, 551, 719
Minnesota Test for Differential Diagnosis of Aphasia, 717
Minnesota Vocational Interest Inventory, 535–536, 719
Minorities, use of tests, 57–64, 191–197, 222–224, 287–298, 343–349, 691–704
Misuse of psychological tests, 30–31
MLA Cooperative Foreign Language Tests, 713
Mode, definition of, 70–72
Moderator variables, 177–180
  and cultural disadvantage, 191–197
  examples of, 178–180
Modern Language Aptitude Test, 423–424, 714
Modifiability of intelligence, 350–354
Mooney Problem Check List, 494–495, 605, 718
Moral Judgment Scale, 554–555
Motor functions, developmental scale, 268–269
  effect of practice, 446
  factor analysis, 445–446
  tests of, 443–446, 715
Multilevel batteries, 305–318
Multiple aptitude batteries, 15–16, 378–388, 711
Multiple choice items, 300, 302, 415–417
Multiple correlation, 182
  shrinkage, 182
Multiple cutoff scores, 184–186
Multiple discriminant function, 190–191
Multiple-factor theories, 371–375
Multiple regression equation, 181–183
  comparison with multiple cutoff method, 185–186
Multitrait-multimethod matrix, 156–158
Musical Aptitude Profile, 458, 716
Musical aptitude tests, 455–458, 716
Myers–Briggs Type Indicator, 518, 718

National anchor norms, 91–92
National Assessment of Educational Progress, 205–206
National Teacher Examinations, 459, 716
Naturalistic observations, 607–608
Needs, measurement of, 510–515, 566, 600
Nelson Biology Test—Revised Edition, 713
Nelson–Denny Reading Test, 713
NLN Achievement Tests in Nursing, 459, 716
Nominating technique, 612–613
Nonlanguage tests, 265, 287, 346–349
Nonnormative scales, 93–94, 96–102
Nonreading Aptitude Test Battery, 387, 711
Nonverbal tests, 265, 346–348
Normal distribution curve, 69–70, 73
  relation to percentiles, 78–80
Normal percentile charts, 79–80
Normalized standard scores, 82–84
Norms, 26, 67, 73–96
  developmental, 73–77
  interrelationships, 88
  local, 92–93
  national anchor, 91–92
  on projective techniques, 578–579
  relativity of, 88–94

  subgroup, 92
  within-group, 77–88

Objective items, 415
Objective personality tests, 588–593
Oblique axes, 369
Occupational interests (*see* Vocational interests)
Occupational tests, 435–463, 715–716
Occupational themes, hexagonal model, 530, 534–535
Odd-even reliability, 115–116
  Rulon formula, 116
Office of Federal Contract Compliance (OFCC), 63, 699–704
Office of Strategic Services (*see* OSS assessment program)
Ohio Vocational Interest Survey (OVIS), 719
Opinion polling, 543–546
Oral Directions Test, 442
Order-of-merit rating, 611
Ordinal scales of development, 76–77, 276–281, 554–555
Ordinal Scales of Psychological Development, 278–279
Organicity, diagnosis of, 467–478
Organizational Climate Index, 619
Orientation of examinees, 36
Orleans–Hanna Algebra Prognosis Test, 423, 714
Orleans–Hanna Geometry Prognosis Test, 423, 714
Ortho-Rater, 715
Orthogonal axes, 369
Orthopedically handicapped, 284–286
Oseretsky Tests of Motor Proficiency, 268–269, 709
OSS assessment program, 19, 595–596
Otis–Lennon Mental Ability Test, 301, 306, 308–309, 710
Otis Self-Administering Tests of Mental Ability, 301, 440, 715
Output, relation to test validity, 172–174
Overlap hypothesis, 328

Pattern analysis, MMPI, 499–504
  Wechsler scales, 465–467
  (*see also* Profiles, examples of)
Peabody Individual Achievement Test, 480
Peabody Picture Vocabulary Test, 285, 709
Peer Nomination Inventory, 613
Peer ratings, 612–613
Percentile bands, 130–131, 315–316
Percentiles, 78–80
  in normal distribution, 78–79
  relation to standard scores, 86–87
Perception tests, as measures of cognitive dysfunction, 471–475
  as measures of personality, 589–591
Performance tests, 265, 281
Persistence tests, 594
Personal Data Sheet, 18
Personal Orientation Inventory, 719
Personality and intelligence, 354–357
Personality inventories, 493–526, 718–719
  content validation, 494–496
  critical evaluation, 524–526
  empirical criterion keying, 496–506
  factor analysis, 506–510
  forced choice, 511 *n*, 517–520
  interpretation of responses, 496–497
  origins of, 18–19
  and personality theory, 510–515
  as self-concept tests, 598–599
Personality Research Form (PRF), 513–515, 718
Personality testing, 18–19, 493–615
  classified list of tests, 718–721
Personnel selection, use of tests in, 3–4, 17, 435–463
    with cultural minorities, 57–64, 191–197, 691–704
Personnel Tests for Industry, 442, 715
Pertinent Questions test, 397, 711
Phenomenology, influence on personality testing, 598–599
Phi coefficient, 213–214
Physically handicapped, tests for, 281–286
Piagetian scales, 77, 276–281, 431–432
Pictorial Study of Values, 719
Pictorial Test of Intelligence, 709
Pimsleur Language Aptitude Test, 424–425, 714
Pimsleur Modern Foreign Language Proficiency Tests, 713
Pintner–Paterson Performance Scale, 281
Piotrowski's Automated Rorschach, 563, 720
Placement, 186–187

Play techniques, as projective tests, 576
Pleasant Events Schedule, 606
Population, IQ changes in, 353–354
  statistical definition, 89–90, 108
Porch Index of Communicative Ability in Children, 717
Porteus Maze Tests, 284, 709
Position Analysis Questionnaire (PAQ), 439
Potentiality, concept of, 25
Power tests, 122
Practice, effect on motor functions, 446
  effect on test performance, 43–44
Preceding activity, effect on test performance, 40–41
Prediction versus diagnosis, 24–25, 140–141
Predictive validity, 140–141
  of infant and preschool tests, 333–334
Preliminary SAT, 318
Preschool testing, 266–271, 273–281, 332–336, 425–434, 573–574, 714
Prescriptive Mathematics Inventory, 714
Prescriptive Reading Inventory, 422, 714
Preselection, effect on validity coefficient, 164
Primary mental abilities, 372–373
  tests of, 378–379
Primary Reading Profiles, 714
Primary school children, group intelligence tests for, 307–309
Primary Visual Motor Test, 717
Privacy, protection of, 49–52
Professional ethics in testing, 45–57, 682–690
Professional school applicants, 459, 460–462
Professions, tests for, 458–463, 716
  certification tests, 459, 462–463, 716
Profiles, examples of, aptitude tests, 130, 316
    interest tests, 529–533
    personality inventories, 499–502, 511
  interpretation, 129–131
  on Wechsler scales, 465–467
Prognostic tests in education, 423–425, 714
Progressive Matrices (Raven), 284, 291–293, 710
Project PLAN, 96
Project TALENT, 91–92
Projective techniques, 19, 558–587, 719–720
  in attitude measurement, 559, 568, 576
  as clinical tools, 586–587
  critical evaluation, 576–587
  examiner variables, 40, 562, 578
  examples, 559–576, 719–720
  nature of, 558–559
  projective hypothesis, 584–585
  as psychometric instruments, 585–586
Psychoeducational Inventory of Basic Learning Abilities, 482, 717
Psychological Screening Inventory, 718
Psychological tests, administration, 32–34
  classified list, 708–721
  definition, 23
  history, 5–19
  and minority groups, 57–64
  misuses, 30–31
  principal characteristics, 23–30
  qualified examiner, 31, 45–48
  social and ethical problems, 45–64
  sources of information about, 20–22
  uses, 3–5
  varieties, 10–19
Psychologists, certification and licensing, 462–463
Public opinion surveys, 544
Pupil Rating Scale (Myklebust), 480, 717
Purdue Pegboard, 444, 715

Q sort, 601, 613–614
Quick Test, 709

Range, 71
  effect on correlation, 125–127, 163–164
Rapport, 34–37
  with projective techniques, 577
Raters, training of, 612
Ratings, 609–612
  behaviorally anchored scales, 609–610
  as criterion measure, 145–146
Ratio IQ, 12, 60–61, 84–85
Raw score, 26, 67–68
Readiness tests, 425–434, 714
Reading tests, diagnostic, 417–420, 713–714
  readiness, 426–429, 711
Reasoning tests, 397, 711

Rectangular distribution, 203 *n*
Reference axes, 364–366
Reference groups, fixed versus normative, 93–94
  in vocational interest tests, 533–534, 538–539
Regression equations, 181–183, 192–196
Reinforcement Survey Schedule, 606
Reliability, 27–28, 103–133
  and ability level, 127, 240
  and age, 240–242
  of criterion-referenced tests, 131–133
  in fan-shaped distributions, 240
  index of, 120–121
  of infant and preschool tests, 262–263
  and item sampling, 113
  projective techniques, 579–580
  relation to test length, 115–116
  and sample heterogeneity, 125–127
  of speeded tests, 122–125
  of Stanford-Binet, 239–242
  types of, 110–122
Reliability coefficient, 109–110
  relation to standard error of measurement, 127–129
Remote Associates Test (RAT), 711
Repeated testing, effect on scores, 43–44
Report writing, in clinical diagnosis, 487–490
Research, use of tests in, 4
Response sets, 515–521
Response styles, 520–521
Retest reliability, 110–112, 120–121
Revised Beta Examination, 710
Revised Minnesota Paper Form Board Test, 447–448, 715
Revised Visual Retention Test, 471–473, 717
Role Construct Repertory Test (Rep Test), 603–604
Role playing, 597–598
Rorschach, 559–563, 720
  computer interpretation, 563
  critical evaluation, 561–563
  examiner variables, 40, 562
  norms for special age groups, 563
Rosenzweig Picture-Frustration Study (P-F Study), 572–573, 720
Rotation of axes in factor analysis, 364–366
Rotter Incomplete Sentences Blank, 571, 720
S-R Inventory of Anxiousness, 523
Sample versus population, 89–90, 108
Sampling, of behavior, 23–24
  of persons, 89–90, 108
  of test content, 113
Scale for Measuring Attitude toward the Church, 547–549
Scales, interval, 200–201
  ordinal, 76–77, 200
Scatter, on age scales, 74
  on Stanford-Binet, 236
  on Wechsler scales, 465–467
Scatter diagram, 104–106
Scholastic Aptitude Test (CEEB), 81–82, 87, 93–94, 283, 301, 318–319, 348–349, 710
School and College Ability Tests (SCAT), 283, 306, 314–318, 710
School readiness, 425–426
School Vision Tester, 715
Scorer reliability, 118–119, 120–121
Scores, types of, 73–96
Screening, 186
Screening tests, for industrial use, 439–442, 715
Screening Tests for Identifying Children with Specific Language Disability (Slingerland), 480, 717
Seashore–Bennett Stenographic Proficiency Test, 716
Seashore Measures of Musical Talents, 455–457, 716
Second-order factors, 369
Security of test content, 30–31, 47–48
Seguin Form Board, 6
Selection, versus classification, 186–187
Selection ratio, 167–171
Self-concept tests, 598–601
Self-control tests, 594–595
Self-Directed Search, 719
Self-report inventories, 18–19, 493–557, 718–719
  as self-concept tests, 598–599
Self-understanding, use of tests in, 4
Semantic Differential, 601–603
Senior Apperception Test, 569, 720
Sensory tests, in early test series, 9–10
  examples, 457, 715
  and intelligence, 8–10
  uses in psychology, 443, 457
Sentence completion, as projective technique, 571
Sequential patterning, 76–77, 152–153

Sequential Tests of Educational Progress (STEP)—Series II, 137–138, 404, 712
Ship Destination Test, 397, 711
Short Employment Tests, 716
Short Tests of Clerical Ability, 716
Significance levels, 108–109
Simulation, in occupational testing, 438–439, 443–444
Situational specificity, 521–524, 595
Situational stress tests, 595–596
Situational tests, 19, 593–598
Situational variables, 39–41
  in projective tests, 511
Sixteen Personality Factor Questionnaire (16PF), 509–510, 718
Skewed distribution curves, 202–203
Skills Monitoring System, 422, 714
Slope bias, 192–195
Social Climate Scales, 620, 721
Social desirability variable, 515–521
Social-learning theory, effect on test construction, 522–524, 616
Social maturity, measurement of, 269–271
Socioeconomic scales, 617
Sources of information about tests, 20–22
Southern California Sensory Integration Tests, 482, 718
Southern California Tests of Fluency, Flexibility, and Elaboration, 711
Southern California Tests of Intellectual Abilities, 711
Spatial aptitude tests, 447–448
Spearman-Brown formula, 115
Special aptitude tests, 13–14, 442–458, 715–716
  relation to intelligence tests, 16, 442–443
  relation to multiple aptitude batteries, 443
Speed tests, in cross-cultural testing, 287–288
  definition, 122
  item analysis of, 217–219
  proportion of variance attributable to speed, 124–125
  reliability of, 122–125
Spiral omnibus tests, 301
Split-half reliability, 114–116, 120–121, 123
  Rulon difference formula, 116
SRA Achievement Series, 404, 712
SRA Clerical Aptitudes, 716
SRA Mechanical Aptitudes, 715
SRA Pictorial Reasoning Test, 710
SRA Primary Mental Abilities, 109 *n*, 124–125, 378–379, 711
SRA Reading and Arithmetic Indexes, 712
SRA Short Test of Educational Ability (STEA), 306, 710
SRA Typing Skills, 716
Stability of IQ, reasons for, 327–329
Standard deviation, 72–73
Standard error, of estimate, 165–166
  of intra-individual score differences, 129–131
  of measurement, 103, 127–129
Standard scores, 80–84
Standardization sample, 26, 89–90
Standardization of testing procedure, 25–26
Standards for Educational and Psychological Tests, 22, 63, 134, 140, 159, 670
Stanford Achievement Test, 16, 99, 403, 404, 405–408, 480, 712
Stanford-Binet, 12, 85, 127, 224–225, 230–245, 327–328, 480, 709
Stanford Diagnostic Arithmetic Test, 420–421, 714
Stanford Diagnostic Reading Test, 418–419, 714
Stanford Early School Achievement Test (SESAT), 404, 408, 428–429, 480, 714
Stanford Modern Mathematics Concepts Test, 713
Stanford Test of Academic Skills (TASK), 403, 404, 408, 712
Stanine, 83–84
State-Trait Anxiety Inventory (STAI), 525, 718
State-Trait Anxiety Inventory for Children (STAIC), 718
Statistical significance, 108–109, 131
Statistics, early development, 8–9
  elementary concepts, 104–110
Sten scale, 83 *n*
Stenography tests, 716
Stereotype accuracy, 581
Stern Activities Index, 619, 721
Stern Environment Indexes, 721
  (*see also* College Characteristics Index)

Strategy, in decision theory, 169, 174–177
Street Gestalt Completion, 589–590
Stromberg Dexterity Test, 715
Strong-Campbell Interest Inventory (SCII), 529–536, 719
Strong Vocational Interest Blank (SVIB), 529–536, 719
Structure-of-intellect model, 374–375, 390–394
STS Junior Inventory, 718
STS Youth Inventory, 495, 718
Study of Values, 552–554, 719
Sub-environment scales, 618–619
Suppressor variable, 183
Survey of Interpersonal Values, 719
Survey of Problematic Situations, 523–524
Survey of School Attitudes, 719
Synthetic validity, 149–151, 438
Szondi Test, 720

T score, 83
Talent, maximizing utilization of, 187–190
Taxonomy of Educational Objectives, 136, 627, 653
Taylor Manifest Anxiety Scale, 505 *n*
Taylor–Russell Tables, 169–171
Teacher-made tests, 209–211, 412–417
Telebinocular, 715
Test anxiety, 37–38
Test bias, 191–197, 222–224
Test ceiling, effect on score distribution, 202–203
Test evaluation, suggested outline, 705–706
Test floor, effect on score distribution, 202–203
Test manual, 21
Test orientation procedures, 36
Test publishers, 21, 47–48, 59, 707
Test sophistication, 44
Test specifications, 136, 413–415
Test-taking attitudes, 34–38, 515–521
Test-taking motivation, 34–38
Testing conditions, control of, 33–34
Testing situation, factors specific to, 39–41, 58–59
Tests, classified list, 708–721
Tests of Academic Promise, 403, 404, 712
Tests of Basic Experiences (TOBE), 429, 714
Tests of General Ability (TOGA), 710
Thematic Apperception Test (TAT), 565–567, 720
  adaptations of, 567–569
  and personality research, 567
Thorndike Dimensions of Temperament, 719
Thurstone-type attitude scales, 546–550
  effect of judges' attitudes, 549–550
  for use in industry, 549
Time-limit tests (*see* Speed tests)
Time sampling, 608
Torrance Tests of Creative Thinking, 394–396, 711
Toy tests, 576
Traits, 14–15, 371–378, 506–510, 521–522
Two-Factor Index of Social Position, 617
Two-factor theory, 369–371
Two-way specification table, 413–414
Typewriting tests, 716
Typing Tests for Business, 716

Underachievement, concept of, 401–402
United States Employment Service (USES), 36, 184–185, 384–387, 451, 711
Uses of psychological tests, 3–5
Utility, in decision theory, 174–175

Validity, 28–30, 134–197
  of battery, 182
  and decision theory, 167–176
  dependence upon sample tested, 163–164
  differential, 190, 383
  incremental, 171–172
  of infant and preschool tests, 333–335
  of items, 206–215
  of personnel tests, 435–439
  of projective techniques, 580–584
  relation to criterion output, 172–174
  shrinkage, in cross validation, 182, 219–221
  of situational tests, 596–597
  of Stanford-Binet, 242–245
  subgroup, 177–180, 192–195
  types of, 134–161

Validity coefficient, 28–29, 163–167
  conditions affecting, 163–165
  variation in, 146–149
Values, assessment of, 552–557, 719
  in decision theory, 174–175
Variability, 70–73
  effect on correlation, 125–127, 163–164
Variance, 72
Vineland Social Maturity Scale, 269–270, 709
  adaptation for the blind, 284
Vision tests, 443, 715
Vocational interests, 527–542
  in measurement of personality traits, 541–543
Vocational Preference Inventory, 719
Vocational testing, 435–463, 715–716

Washington University Sentence Completion Test (WUSCT), 571, 599
Watson-Glaser Critical Thinking Appraisal, 397, 711
Wechsler Adult Intelligence Scale (WAIS), 245–255, 709
  abbreviated scales, 249
  age decrement, 337–338
  brain damage, diagnosis, 467–470
  deterioration index, 465
  factorial analysis, 254–255
  pattern analysis, 465–467
Wechsler-Bellevue Intelligence Scale, 246–247
Wechsler Intelligence Scale for Children (WISC-R), 255–260, 480, 708
  factorial analysis, 260
Wechsler Preschool and Primary Scale of Intelligence (WPPSI), 260–264, 480, 710
  factorial analysis, 264
Wesman Personnel Classification Test, 441, 715
Wide Range Achievement Test (Jastak), 480, 712
Wideband procedures, 586–587
Wing Standardized Tests of Musical Intelligence, 457–458, 716
Wonderlic Personnel Test, 440–441, 715
Woodworth Personal Data Sheet, 18, 494, 496
Word association tests, 18, 569–571
Work Environment Preference Schedule, 719
Work Values Inventory, 554, 719

z scores (*see* Standard scores)

30298